Expression-Based Modeling of Metabolic Flux in Metabolic Diseases

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by Yiping Wang
December 2019

Biographical Sketch

The author was born in Nanjing, China, on May 23, 1991. He remembers nothing of the first four years of his life, spent mostly in the nearby town of Anqing, China. From the memories of his parents, though, he knows that he was cherished in the heart of his extended family, and one of his grandfather's favorite grandchildren. The author is a proud American, but retains a deep love for Chinese culture as embodied in his family.

In 1995, the author's parents made the momentous decision to move themselves and their son to Morgantown, West Virginia, United States, to pursue Ph.D.'s at West Virginia University. The author rapidly lost most of his knowledge of Chinese, but gained the knowledge of English, and with it access to his favorite place in the world, the library. Here he found his oldest and truest friends in the pages of novels, and the wonders of the wide world in science, technology, and history books. From there, a dream was kindled in the author that one day he might contribute to the marvels described to him.

In combination with his mother, who pounded into him a rigorous math and computer science education, the author excelled throughout his pre-college academic career in West Virginia. He then enrolled at Duke University, selecting Computer Science as his first major. But the author always wanted to focus on scientific applications of computational techniques, and in his first year, hit upon computational biology as the most challenging, and worthwhile. To this end, he added Biology as his second major, and partic-

ipated for three summers in the Vertically Integrated Fellows program, in undergraduate research. From there, he eventually joined Dr. Pelin Volkan's lab as an undergraduate researcher, studying DNA regulatory motifs that govern the differentiation of the Drosophila olfactory neurons.

After graduation, the author came to Cornell University as a bright and eager first-year in the Tri-Institutional Computational Biology and Medicine Program. Beginning his rotations in Dr. Zhenglong Gu's lab in Ithaca, he found a firm friend in Brandon Barker there. He undertook a dizzying number of classes, rotations, and research projects in his first year and second year, but gradually settled into Dr. Gu's lab, where for several years now he has pursued a project to apply computational modeling of metabolism to cancer and gut microbiome data. The conclusions of that work are presented here in this thesis.

Acknowledgements

So many people have contributed to my growth as a researcher during my time at Cornell. More than any other period of my life, my experiences here have made me an adult. I cannot specifically thank all of the people I would like to here, but nevertheless they will always influence and enrich my life.

Among all of my connections, I want to first thank Brandon Barker for being a true friend to me the whole time. Brandon was finishing up his Ph.D. when I came, and I was very fortunate to overlap with him for a few years. He introduced me to the field of flux balance analysis, and in particular helped me become familiar with the FALCON algorithm. FALCON is his own work, but I am proud to have assisted at its creation, and it has served as the foundation of my subsequent research. Furthermore, Brandon introduced me to graduate life in general, its tribulations and its joys. Even after he graduated, we have gotten together sometimes for lunch, and to continue talking about research.

My committee members have each contributed significantly to my research direction. Dr. Olivier Elemento has always given me good ideas when he videoconferenced at each of my yearly committee meetings. He has also given me great advice about my career after my Ph.D. Likewise, Dr. Haiyuan Yu has also contributed significantly each year that I met him, and helped me to put more statistical rigor and clarity in my work. Dr. Chris Myers hosted me in his lab for one rotation during my first year, which helped me to develop greater expertise in metabolic modeling. Over the years, I have

also met with him several times to discuss my work, and each time he has brought a critical and rigorous perspective to my work. Finally, I want to thank my committee chair, Dr. Zhenglong Gu. Zhenglong believed in me at multiple points during my Ph.D., when noone else could, and gave me the crucial support I needed to continue. We have had many discussions over the years about all of my work, and he has always pushed me both to expand the boundaries of my creativity, and also examine my work with rigor. If I have improved at all as a researcher during the last several years, it is due to him.

My most important thanks go to my family, my girlfriend, and to God. My family raised me with love all my life, and set me on the path to obtaining a Ph.D. I have relied on the rock of their love throughout my time at Cornell, and always will. They prepared me for something I never thought would happen: to find a woman who I would love, and who loved me back. The greatest miracle is that I found such a beautiful soul as Shaohuan Wu, the light of my life, and we came together in love for each other. And finally, I want to thank God who created the world, and all its miracles, and me. For so long I did not know him, but at Cornell finally found faith, and recognized him. All things are in his hands, and may he always guide and protect me, and all his children.

Expression-Based Modeling of Metabolic Flux in Metabolic Diseases

Yiping Wang, Ph.D. Cornell University 2019

Metabolism is one of the most central aspects of the biology of all organisms. All cells possess a bipartite metabolic network, composed of small molecule metabolites and enzymes which carry out biochemical reactions on them. This metabolic network is responsible for carrying out numerous cellular functions, including energy generation in the form of ATP, production of antioxidants to control reactive oxygen species levels, and production of biosynthetic molecules necessary for cellular growth, such as amino acids and nucleic acids. However, in Chapter 1, I will describe how one of the most important facets of metabolism, the rate at which each biochemical reaction occurs, or metabolic flux, cannot be easily experimentally measured on a genome-wide scale. This leads us to the need for a computational method that can efficiently infer such metabolic flux. I will describe the approach taken by a group of methods that go by the general name of constraints-based modeling, and how we have previously developed a new method under this framework, called FALCON, which uses metabolic gene expression to predict flux. I will then describe how I applied FALCON to infer differences in metabolic flux in two major groups of metabolic diseases.

First, metagenomic sequencing has revealed that the composition of the gut microbiome is linked to several major metabolic diseases, including obesity, type 2 diabetes (T2D), and inflammatory bowel disease (IBD). I used

the computational tool PICRUST to infer species-specific metagenomes for each of these diseases, and FALCON to infer fluxes from these results. I discovered that several major pathways, previously shown to be important in human host metabolism, have significantly different flux between the two groups. I also modeled metabolic cooperation and competition between pairs of species in the microbiome, used this to determine the compositional stability of the microbiome, and found that that the microbiome is generally unstable across controls as well as metabolic microbiomes.

Second, I also used RNA-Seq data from The Cancer Genome Atlas (TCGA) as input to FALCON. I found a systematic difference in that cancer tissues have a considerably stronger correlation between RNA-seq expression and inferred metabolic flux, which may indicate a more streamlined and efficient metabolism. I also found several pathways that frequently have divergent flux. Among these are sphingolipid metabolism, methionine and cysteine synthesis, and bile acid transformations.

Contents

Chapter 1

Alterations of Metabolic Flux in Metabolic Diseases

1.1 A Short History of Cancer Metabolism

Over the past few decades, the field of cancer metabolism has undergone a renaissance [1]. Its roots go back at least to the 1920's, when Otto Warburg made the first seminal observations of increased glucose uptake and fermentation in cancer cell cultures, even under normoxic conditions that normally suppress fermentation [2]. These observations were later complemented by the discovery of the first chemotherapeutic drugs, such as 5-FU in the 1950s [3]. The molecular mechanisms behind many of these lay in their ability to restrict one-carbon metabolism, which is preferentially used by proliferating cells such as cancer [3]. Nevertheless, the degree of interest in this field later declined, as the genetic basis of cancer with oncogenes and tumor suppressors was discovered [1]. Initially, most of these seemed to have little connection with cancer metabolism citecai.

However, several related developments have recently led to a resurgence of interest in the field. Many new metabolic pathways, such as glutaminolysis [4] and fatty acid synthesis [5], have been discovered to have a major role in tumorigenesis. It turned out that some classic oncogenes and tumor suppressors, such as p53 [1], have unexpected roles as metabolic regulators. And crucially, it is now believed that cancer metabolic and signaling networks are interlinked through several cases of reciprocal regulation [6]. Signaling networks can affect the mRNA expression, protein translation, and degradation rates of metabolic enzymes [7]. Metabolic networks can control the levels of small-molecule metabolites such as acetyl-CoA and various forms of folate [8]. These metabolites can then act as substrates for the creation of epigenetic marks, as well as bind to signaling proteins and alter their activity [9].

Together, these results have led to concept that cancer can be considered a metabolic disease, in the same way as diabetes and obesity [10]. This is supported by the high rates of cooccurrence among all of these three diseases [11].

1.2 The Metabolic Needs of Cancer Cells

Cancer cells are defined by their uncontrolled proliferation, at rates and in locations that are outside the bounds of normal body cells [12]. In turn, such proliferation requires many inputs from cell metabolism in order to occur. ATP for energy demands is only one of these, and often not the limiting

factor. Proliferation also requires great amounts of nucleotides, amino acids, and lipids, often in relatively greater amounts than ATP [13]. Synthesis of these compounds requires, among other things, large amounts of NADPH and acetyl-CoA. A possible explanation for the Warburg effect is to quickly supply these needs [13].

Large glucose fermentation has a "by-product" effect, in that it also allows large production of NADPH, through the oxidative pentose phosphate pathway (PPP), as well as acetyl-CoA through the pyruvate dehydrogenase complex [1]. To prevent glucose from entering oxidative phosphorylation as opposed to fermentation, cancer cells express the pyruvate kinase M2 (PKM2) isoform, which catalyzes the last step of glycolysis, instead of PKM1 which is found in most normal cells [14]. PKM2 is a much slower enzyme than PKM1, which however allows the concentration of upstream glycolytic intermediates to build up [1]. These intermediates are thus forced to enter the PPP, which is an alternative route for their metabolism. NADPH production in particularly crucial for cancer cells, as they generally have high levels of ROS [1]. NADPH acts as a substrate for generation of the antioxidant molecules glutathione and thioredoxin, which help to control the deleterious effects of ROS [1]. Additionally, glutamine can also act as an additional source of acetyl-CoA, in the pathways of glutaminolysis and reductive carboxylation (involving a short part of the TCA cycle operating in reverse), eventually resulting in conversion to acetyl-CoA [4].

1.3 Control of Cancer Metabolism by the Cancer Signaling Network

Several major signaling pathways have been found to have a major effect on reprogramming cancer cell metabolism in order to support increased proliferation. One of the most important of these is the phosphatidyl 3-kinase (PI3K) pathway. PI3K enzymes at the cell membrane are activated by receptor tyrosine kinase (RTK) signaling, and phosphorylate inositol in the cell membrane at the 3' position [15]. The PTEN enzyme carries out the opposite effect, dephosphorylating PI3 to inositol. In cancer, activating mutations in PI3K or inactivating mutations in PTEN lead to increased levels of PI3, which in turn activates the intracellular kinase Akt [16]. Akt has been shown to stimulate transcription of many metabolic enzymes, including glucose transporters, hexokinase (HK), and phosphofructokinase 2 (PFK2), all of which are components of the glycolytic pathway [1]. Akt also activates the mTOR complex, which upregulates the rate of ribosome biogenesis and mRNA translation, leading to increased protein biosynthesis [17].

mTOR also activates the expression of the hypoxia-inducible factor (HIF) complexes, HIF1 and HIF2 [17]. HIF1 is composed of the two subunits HIF1b and HIF1a, and HIF2 of HIF1b and HIF2a [18]. Under normoxic conditions, the HIF complexes are recognized and marked by prolyl hydroxylase enzymes, leading to their degradation by the E3 ubiquitin ligase von Hippel-Lindau factor (VHL) [18]. Under conditions of hypoxia, which is common in the

tumor microenvironment, or mutations in VHL, HIF acts as a transcription factor that further upregulates expression of many glycolytic enzymes and transporters [18].

Finally, the well-known tumour suppressor p53 has also been shown to have major effects on cellular metabolism. p53 does upregulate the expression of HK, but also upregulates TIGAR, an enzyme that decreases levels of the glycolytic allosteric activator fructose-2,6-bisphosphate [19]. p53 also promotes oxidative phosphorylation as opposed to fermentation, by activating expression SCO, a factor required for assembly of the cytochrome c complex in the electron transport chain [19]. Finally, p53, through its activation of p21, also supports the expression of the transcription factor NRF [20]. NRF is considered the master antioxidant regulator in the cell, whose activity is required to control levels of ROS which often accumulate during tumorigenesis.

1.4 Control of Cancer Signaling by the Cancer Metabolic Network

Several major classes of histone modifications, such as acetylation, methylation, and N-Acetylglucosamination (GlcNAcylation), depend on "writer" enzymes [9]. These enzymes add or remove post-translational modifications to histones, using as substrates small molecules that are produced by the metabolic network. Acetylation is performed by acetyltransferases, using

the molecule acetyl-CoA, which can be synthesized by any of the three enzymes ATP-citrate lyase, pyruvate dehydrogenase complex, or acyl-CoA synthetase. Methylation relies on methyltransferases, which use the metabolite S-Adenosylmethionine (SAM), synthesized by the enzyme methionine adenosyltransferase (MAT) as part of the one-carbon cycle [21]. In addition, demethylation is performed by demethylases of the Jumonji-C or TET families, both of which rely on the TCA cycle metabolite alpha-ketoglutarate as a substrate [21]. Finally, GlcNAcylation is carried out by O-GlcNac transferase (OGT), using UDP-glucosamine produced by the hexosamine synthesis pathway. The inputs to hexosamine synthesis are ultimately either glucose or glutamine [9].

Changes in small metabolite substrate concentrations may potentially alter the activity of writer enzymes significantly, affecting the chromatin state of hundreds of genes, including some which are components of signaling pathways [9]. Furthermore, it has recently been shown that cancer cells may produce a novel metabolite called 2-hydroxyglutarate (2-HG) [22]. 2-HG is formed the enzyme isocitrate dehydrogenase, which normally produces alpha-ketoglutarate, if a mutation occurs there [22]. Along with fumarate and succinate [23], 2-HG has been shown to be a potent inhibitor of Jumonji-C and TET demethylases, which normally use the structurally related alpha-ketoglutarate as a substrate.

Together, these examples show changes in metabolite concentrations may act upon cancer signaling networks. This route of signaling regulation is

particularly important, because metabolite concentrations are directly linked to the state of the metabolic network, and hence to the nutritional and metabolic environment that a cell is located in [9].

1.5 Structure and Diversity of the Human Gut Microbiome

Recent advances in environmental shotgun sequencing have led to an enormous increase in our knowledge of the human gut microbiome [24]. 16s sequencing focuses only short variable regions of the 16s rRNA gene of bacteria, leading to a catalogue of species present in a sample [25]. Conversely, metagenomic and metatranscriptomic sequencing allow measurement of the abundance and expression of all genes in a sample, though generally without information on which species are represented [26]. Together, these two methods have allowed the description of the structure and diversity of the human gut microbiome, which in healthy individuals is composed of approximately 1000 operational taxonomic units (OTUs) [24]. OTUs are groups of closely related 16s sequences, which are generally considered to represent separate microbiome species [24].

Assembly of the gut microbiota begins after birth, from microbes initially present in the maternal vagina or skin, and gradually increases in size and diversity over development [27]. In an adult human, the gut microbiota is now estimated to contain 10 times as many individual cells as the human host,

and 100 times as many genes [28]. Nearly all of them are of bacterial origin, although archaeal, fungal and viral species have also been detected [28]. The highest numbers are found in the colon of the large intestine, at levels that make it the most densely populated bacterial environment in the world [29]. Two major bacterial phyla, Bacteroides and Firmicutes, dominate this location, although other phyla like Proteobacteria and Actinobacteria are also present [29]. Bacteria in this location face warm, anaerobic conditions, with generally stable amounts of food from both dietary sources, as well as mucus secreted by intestinal cells [30]. Nevertheless, the types of food available from the diet can shift quickly, leading to rapid changes in microbiome composition [31].

Despite this, large-scale studies of the microbiomes of healthy individuals, such as the Human Microbiome Project (HMP), have shown that the within-individual species composition of each person, based on 16s results, is stable over time [25]. However, when comparing different individuals in a population, major differences in their composition emerge, both at the level of phyla such as Bacteroides vs. Firmicutes, as well as individual species [25]. Furthermore, metagenomic studies of the same individuals reveal that the overall abundance of metabolic pathways is stable, even across individuals with different species compositions [25]. Under disease conditions, 16s and metagenomic studies have also shown that microbiota composition and function can change significantly in disease conditions, such as diabetes [32] and inflammatory bowel disease (IBD) [33].

1.6 Effects of Gut Microbiome Metabolism on Human Health

The gut microbiota carries out three major functions for its human host. First, the large number of species causes intense competitive pressure for space and nutrients, which helps prevent pathogenic bacteria from potentially gaining a foothold [28]. Second, interaction with the gut microbiota, such as through exposure to antigens, is required to help guide the proper development of some host tissues, including intestinal epithelial cells and the innate immune systems [28]. Finally, the gut microbiome encodes a vast number of metabolic enzymes, which can cooperate with the host metabolism to provide otherwise inaccessible nutrients and energy [28].

Gut bacteria have the ability to synthesize essential vitamins and cofactors, which the host can then absorb [30]. However, the most intensively studied aspect of microbiota metabolism has been their ability to break down inaccessible complex polysaccharides, or glycans [30]. This large and chemically diverse class of compounds is found in plants (starch, hemicellulose and pectin), animals (glycosaminoglycans and N-linked glycans), and host intestinal mucus secretions (O-linked glycans) [30]. Humans have enzymes that are only capable of digesting a few of the most common glycans. The remainder must be digested by the microbiome, which collectively encodes hundreds of the glycoside hydrolase enzymes needed to break glycans into individual sugar monomers [30]. These monomers are then typically fermented

into short-chain fatty acids (SCFAs) such as butyrate, acetate and propionate. SCFAs are then released into the intestinal lumen, and absorbed by the intestinal lining. Butyrate is an important energy source for colonocytes, which use it as a substrate for the TCA cycle, ans higher butyrate usage is associated with lower risk of colon cancer [29]. Acetate and propionate enter the bloodstream and are later metabolized in the liver. They can also serve as signaling molecules which play an important in gut innate immunity [29].

Finally, humans are exposed to a huge variety of dietary compounds, environmental chemicals, and drugs that are described as xenobiotics. These foreign compounds are typically transformed by host enzymes, such as the cytochrome P450 system, into more water-soluble forms, which can then be safely excreted [34]. However, the gut microbiome is potentially a huge source of metabolic activities that may alter xenobiotic structure independently of the host, with results that are either beneficial or deleterious for the host [34]. Analysis of gut metagenomes already shows that they contain many enzymes that may carry out such activity, though further work is needed to confirm their function. At the present time though, three separate studies have already shown that the gut microbiome may metabolize the drugs digoxin and irinotecan, as well as the dietary compound choline into trimethylamine N-oxide (TMAO). Digoxin, a treatment for heart attacks whose dosage must be carefully controlled, is transformed into the ineffective form dihydroxydigoxin [35]. Conversely, irinotecan is an effective cancer drug whose deleterious side effects may be reduced by liver glucuronidation.

However, beta-glucuronidases in the gut microbiome may remove the glucuronidation, turning irinotecan into its original form, which causes intestinal damage [36]. Finally, TMAO formed from choline has been implicated in prevalence of several diseases, including cardiovascular disease [37].

Chapter 2

Modeling of Metabolic Flux Differences in Metabolic Diseases

2.1 Constraint-Based Modeling of Metabolism

Constraints-based models (CBM) are one of the most widely used and successful ways of modeling metabolism [38], [39]. They are based on the simple equation S*v = dx/dt. Here, v is a vector of fluxes in the metabolic network that are to be calculated, dx/dt is a vector of rates of change of all metabolites in the network, and S is the stoichiometric matrix. S has one row for each metabolite in the network, one column for each reaction, and a stoichiometric coefficient at each row-column intersection, which determines how many molecules of a metabolite are produced or consumed in each reaction [40].

The goal of CBM is to predict the vector v, based on two key assumptions about metabolism. First, the steady-state assumption states that the

concentrations of each metabolite should settle to a steady-state value, so that S*v=dx/dt=0. This makes intuitive sense, as concentrations cannot indefinitely rise or fall under any physical situation. Second, the rate constants and enzyme expression for any reaction are finite, so the total flux through any reaction must be bounded. Although the exact limit is unknown for most reactions, a very permissive upper bound of 1000 flux units is typically used, as well as a lower limit of -1000.

These two assumptions define a feasible subspace of flux distributions, which is shaped as a convex cone. Most CBM methods, like flux balance analysis (FBA) then find the flux distribution within this subspace that maximizes biomass production [39]. These methods assume that cells regulate their flux distributions to grow as quickly as possible, which has been validated experimentally for bacteria in several papers [41], [42], [43]. However, most tissues in multicellular organisms are not actively growing despite sufficient nutrients to do so. Rather, diverse cell types, like neurons, muscle cells, and kidney cells in humans, regulate their metabolism to carry out tissue-specific functions [44]. Furthermore, in the case of the gut microbiome, it is known that individual species in the microbiome may try to optimize objectives very different from biomass, such as synthesis of antimicrobial compounds against competitors.

2.2 The FALCON Algorithm, a Continuous Gene Expression CBM Method

Several approaches have been proposed for defining a multicellular flux objective. Certain tissues like liver and kidney are specialized for exchanging metabolites with the bloodstream. Maximizing total flux through these exchange reactions has been shown to give good agreement with experimental measurements [45], [46]. However, this approach fails for non-secretory tissues like muscle and adipose. A much more general, unbiased, and widely used objective is to maximize the correlation between predicted flux and enzyme expression [47]. This relies on the idea that high-flux reactions require high enzyme levels to catalyze them, while low-flux reactions would tend to have low enzyme levels to avoid the cost of synthesizing unused enzymes [48]. Several computational algorithms have been published using this approach, which fall into two major classes, discrete and continuous.

Discrete algorithms set an expression threshold for each reaction, beneath which reactions are considered inactive because they do not have enough catalyzing enzyme. These reactions are removed from the metabolic network, and then FBA is run on this reduced model [49] [50]. Most discrete methods, like GIMME [51] and MBA [46], also allow lowly-expressed reactions to be added back if it turns out they are necessary to create biomass [48]. However, it has been shown that low-expressed reactions may carry small fluxes which are nevertheless essential to cellular function, yet are not related to creating

biomass. Continuous methods do not use an expression threshold and thus avoid this problem. Instead, the optimization parameters for each reaction are continuously varied as a function of expression. For example, E-Flux modifies the upper bound of each reaction, so that it is linearly proportional to the enzyme expression [52].

We have implemented FALCON, a continuous method which takes a more direct approach than E-Flux. FALCON implicitly maximizes the linear correlation between flux and expression, by minimizing the objective $S = \sum_i \frac{v_i - d_i}{\sigma_i}$, where i runs over all reactions in a metabolic network, v_i stands for the flux of the ith reaction, d_i stands for average gene expression associated with the ith reaction, and σ_i stands for the standard deviation of gene expression measurements [65]. Furthermore, FALCON offers improved handling of gene-protein-reaction rules. All gene-expression-based CBM methods rely on these rules, which map gene ids to their reactions in a metabolic network. Multiple genes may map to a single reaction, as in the case of enzyme complexes or isozymes. In these cases, we have shown that different logically equivalent ways of writing a GPR rule may give different values for the overall expression associated with a reaction [53]. FALCON solves this issue by converting all rules to a consistent conjunctive normal form (CNF) format before assigning expression.

2.3 Application of FALCON to Gut Microbiome Data

Using constraints-based methods to predict microbiome metabolism presents special problems as well as opportunities. A natural framework is to regard each microbial species as a separate compartmented metabolic network, exchanging metabolites with a common extracellular compartment representing the environment [54]. Host metabolic inputs and outputs can be defined to this environment, which may be correlated with known disease and healthy metabolic states.

Nevertheless, two key problems prevent the full potential of CBM methods from being used to model the gut microbiota. First, it is unclear what optimization criterion should be used when modeling microbiota metabolism [55]. Maximizing biomass production with FBA is not viable, as this would predict that a single most efficient species would outcompete all others in the microbiota, which is clearly not observed experimentally [56]. An alternative is to assume that some species metabolically cooperate with each other to maximize their biomass [56], [57], but this requires assuming cooperation a priori.

Several previous methods have been proposed for modeling microbiome metabolism in a constraints-based framework. Three of the most prominent ones are CASINO [58], SteadyCom [59], and MAMBO [60]. MAMBO using data on relative abundance of individual species in metagenomic samples.

Initially, metabolic models for each of these species are assumed to have access to a uniform set of nutrients in the environment. MAMBO then uses flux balance analysis, coupled with a Monte Carlo algorithm, to calculate an individualized metabolic environment for each species, such that it optimizes the correlation of species's FBA growth rates with their measured abundances. SteadyCom, on the other hand, is quite similar to traditional FBA, except that it imposes a uniform growth rate on all species in the microbiome, and then optimizes this uniform growth rate. And CASINO uses both individual species biomass objectives as well as a combined community objectives, iteratively optimizing both of them to obtain a solution.

Despite the development of these and other methods, no method has yet been applied to model all possible species in the gut microbiota as a single community [61]. This allows an opportunity to apply FALCON as the first CBM method that analyzes the entire gut microbiota.

2.4 Benchmarking of Assumption About Uniform Protein Complex Stoichiometry

As noted above, FALCON converts the GPR rules of any model to CNF, and uses this consistent form to calculate the expression levels of enzyme complexes and isozymes from expression levels of individual genes. FALCON also makes a key assumption regarding the stoichiometry of each enzyme complex. Each enzyme complex A may be composed of individual subunits

A1-An , and in general different numbers of each subunit may be required for a functional complex. For example, the F1 subcomplex of ATP synthase requires one copy of the subunit, and three copies each of the and subunits [62]. However, enzyme complex stoichiometry is not known for most complexes, and not included in metabolic network models. Therefore, FALCON assumes that all complexes have uniform stoichiometry, that is only one copy of each known subunit is required for a functional complex..

To our knowledge, the only metabolic model that does include enzyme complex stoichiometry is the Whole Cell model of Mycoplasma genitalium, a small human parasite by Karr et al. [63]. Indeed, the Whole Cell model integrates all known experimental information about M. genitalium, and can be used for stochastic simulation of all molecular interactions in the organism, such as transcription, translation, and DNA replication. For metabolism, the Whole Cell model uses a dynamic FBA method [64]. Briefly, at each time step t, the metabolite concentrations are used to constrain which fluxes are active in an FBA calculation, and then the FBA fluxes are used to update the metabolite concentrations.

We therefore used the Whole Cell Model to investigate the possible effect of uniform versus experimental enzyme complex stoichiometry. We created a custom modification of the Whole Cell Model's MATLAB source code that used uniform instead of experimental stoichiometry. We then ran either the original Whole Cell Model, or our modified version, for nine hours, simulated growth of a single M. genitalium cell, and plotted the overall difference in

metabolites and fluxes between the two conditions (2.1). Please refer to the legend of 2.1 for details.

Remarkably, when we plotted either the total amount of flux in each reaction, or concentration of each metabolite, in one setting versus the other, we observed a near perfect log-linear correlation. Albeit there are a few outliers, these results show that uniform versus experimental stoichiometry does not make a great difference in flux predictions under the comprehensive Whole Cell model of all cellular processes. We can therefore predict that there should not be a great effect in FALCON simulations, which involve only metabolism. Our results have been published in Computational Biology and Chemistry [65].

2.5 Benchmarking on Accuracy for Predicting CORE profiles of NCI-60 cell lines

The gold standard for validation of any CBM method is comparison of its flux predictions to experimental flux measurements. Although we were unable to find a comprehensive dataset measuring intracellular eukaryotic fluxes, we used a very good substitute in the CORE profiles of the NCI-60 cell lines, as measured by Jain et al. [66] This dataset was obtained by growing each of the cell lines in RPMI-1640 medium, and measuring metabolite concentrations by mass spectrometry in fresh and spent medium. The CORE profile thus represent the estimated total uptake/release of each metabolite by a cell line.

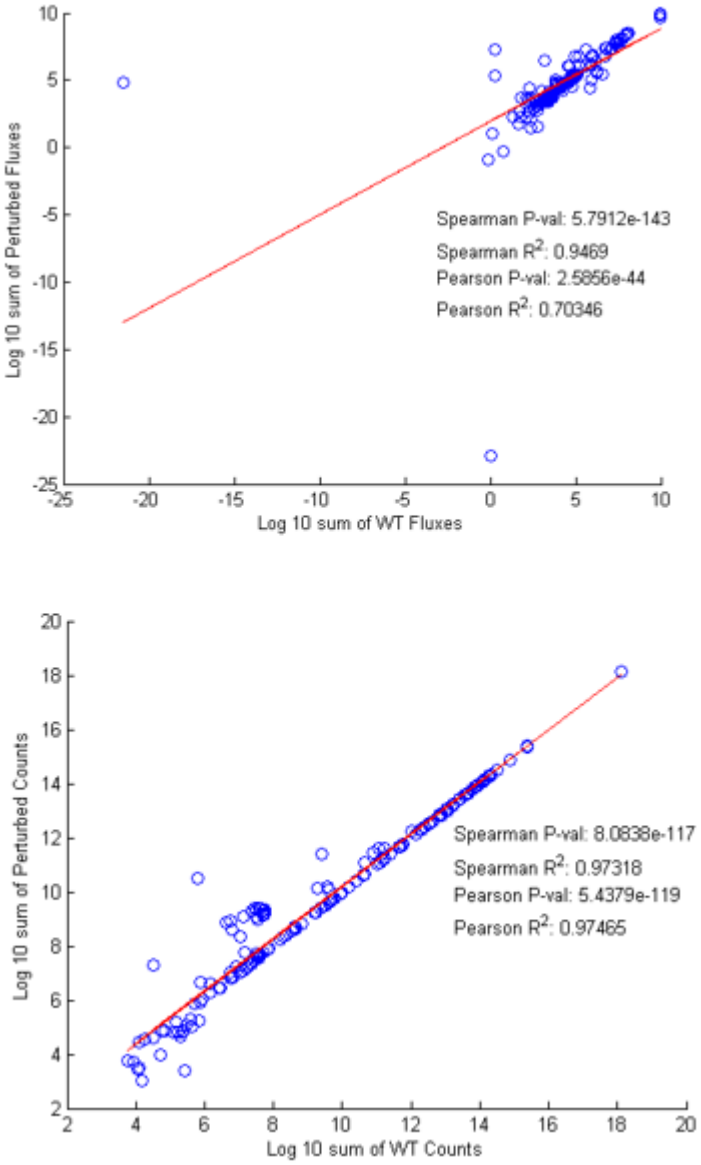


Figure 2.1: Above: The Whole Cell model was simulated for 9000 seconds with both default (WT) and simplified unitary (Perturb) protein complex stoichiometry. Among 645 total metabolic fluxes, 298 had nonzero flux in both conditions. We calculated the sum of each of these 298 fluxes across 9000 seconds. We omitted 10 fluxes whose overall sum was of different sign between WT and Perturb, took the log10 value of the remaining 288 fluxes, and plotted these values, in both WT and Perturb (top). Similarly, among 722 total metabolites, 182 had nonzero counts in both conditions. We calculated the sum of each of the counts of these 182 metabolites across 9000 seconds, took the log10 of these values, and plotted them in both WT and Perturb (bottom). In both cases, the red line indicates the line of best fit for the Pearsons correlation.

Our goal then was to perform a comprehensive comparison among FALCON, FBA, and a panel of other gene-expression CBM methods, on their ability to predict the direction of uptake/release as determined from the CORE profiles. Although these methods have been compared on bacterial datasets [67], their performance on multicellular human expression data has never been thoroughly tested to our knowledge. For these simulations, we used either Recon 2, a comprehensive model of all reactions in human metabolism, or submodels as described below. We also obtained expression data, if necessary, from proteomics on the NCI-60 cell lines in RPMI-1640 medium performed by Gholami et al. [68].

Details of each algorithm we ran are listed below: Normal FALCON: We used the full Recon 2 model, and FALCON with proteomics data from Gholami et al. Normal FBA: We used the full Recon 2 model, and applied FBA using the generic biomass reaction, and the function optimizeCbModel from the COBRA Toolbox [69]. GIMME and iMAT FALCON and FBA: same as for Normal FALCON and FBA, except that we first applied a discrete gene-expression CBM method to obtain a submodel of Recon 2. We wished to see whether restricting FALCON or FBA to only high-expression reactions would lead to more accurate flux predictions. The implementations used here come from the COBRA toolbox [69]. GIMME and iMAT Machado FALCON and FBA: same as above, using an alternative implementations of GIMME and iMAT by Machado et al. [67] that were reported to be closer to the original published versions. EFlux: We used the full Recon 2 model,

and applied EFlux using the Gholami expression data. GXFBA: We used the full Recon 2 model, and applied GXFBA using the Gholami proteomics data. Additionally, GXFBA requires a reference flux distribution as a base for predicting cell line-specific fluxes. We used the Normal FBA solution as an unbiased reference.

We calculated prediction accuracy only for 91 metabolites from the CORE data that had corresponding exchange reactions in Recon 2. Furthermore, it was shown through FVA [70] that Recon 2 was not capable of performing uptake/release of all of these 91 metabolites in the CORE dataset. Therefore, for each metabolite and cell line, we first check whether the FVA analysis allows exchange flux that matches the CORE data. If so, we then record correct prediction if the direction of predicted flux matches the CORE data, and otherwise (including 0 predicted flux) an incorrect prediction. Accuracy is calculated as the percentage of correct predictions for each cell line.

All four of the Machado GIMME or iMAT settings removed nearly all corresponding exchange reactions of the CORE metabolites. As a result, their accuracies are all zero or close to it, as shown in 2.2. Also, FALCON actually does not do as well in FBA in overall accuracy wrt to predicting CORE fluxes, whether the full Recon 2 model is used ("Normal"), or tissue-specific models created using two different algorithms ("iMAT" and "GIMME"). Furthermore, both of the other two methods for predicting flux from prediction (EFlux and GXFBA) do somewhat better than both FALCON and FBA. This is not too surprising, though, since both of these methods actually in-

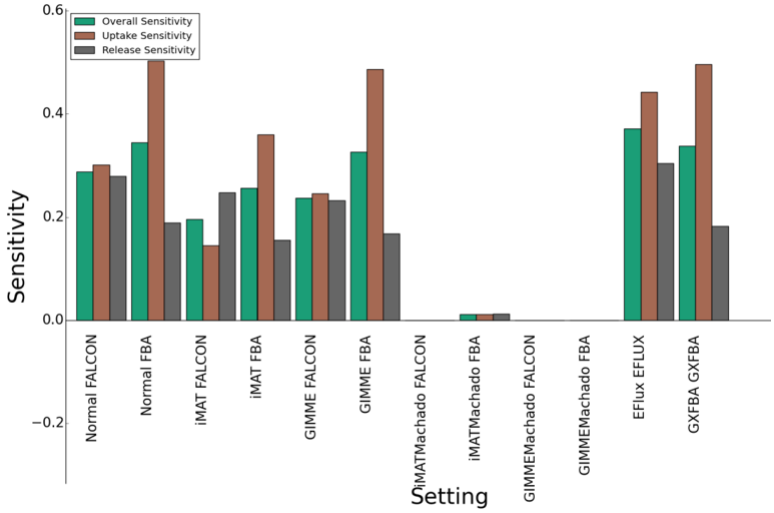


Figure 2.2: Above: Average accuracy across all cell lines was calculated for all uptaken metabolites in the CORE data (brown), released (gray), or all metabolites overall (green).

corporate FBA as part of their overall pipeline.

Chapter 3

Expression-Based Inference of Metabolic Flux Differences in Human Cancer

3.1 Abstract

Cancer cells display numerous differences in metabolic regulation and flux distribution from noncancerous cells, which are necessary to support increased cancer cell growth. However, current experimental methods cannot accurately measure such metabolic flux differences genome-wide. To address this shortcoming, we apply FALCON, a computational algorithm for inferring metabolic fluxes from gene expression data, to analyze data from The Cancer Genome Atlas (TCGA). We found several major differences between tumor and control tissue metabolism. Cancer tissues have a considerably stronger correlation between RNA-seq expression and inferred metabolic flux, which may indicate a more streamlined and efficient use of metabolism. Cancer

metabolic fluxes generally have high correlation with their normal control counterparts in the same tissue, but surprisingly, there are several cases where tumor samples in one tissue have even higher correlation with control samples in another tissue. Finally, we found several pathways that frequently have divergent flux between tumor and control samples. Among these are several previously implicated in tumorigenesis, including sphingolipid metabolism, methionine and cysteine synthesis, and bile acid transformations. Together, these findings show how cancer metabolism differs from normal tissues and may be targeted in order to control cancer progression.

3.2 Introduction

Alterations in metabolism have recently been identified as one of the hallmarks of cancer [12]. This recognition comes in the wake of numerous studies showing that cancer cell metabolism is broadly dysregulated [1]. Indeed, many common driver mutations have been shown to cause coordinated changes in cancer signaling and metabolic networks, reprogramming their metabolism in order to support the demands of continued proliferation [1]. In normal cells, proliferation takes place when growth factors bind to receptors such as IGFR, EGFR, and PDGFR, leading to upregulation of pro-growth signaling pathways such as PI3K-Akt, mTOR, and c-Myc [17] [71] [72]. Mutations in either these receptors, or components of the downstream signaling pathway proteins, can lead to changes in their activity patterns in cancer

cells. In turn, this leads to changes in metabolic regulation, including but not limited to changed expression and degradation rates of metabolic enzymes and transporters, different patterns of phosphorylation and acetylation, and altered concentrations of allosteric regulators such as fructose-2,6-bisphosphate [1].

These regulatory changes are believed to cause a rewiring of metabolism in cancer, so that flux through metabolic pathways is reorganized in order to support increased cell proliferation [13]. The oldest and most well-studied of these pathway changes is the Warburg effect, first observed by Otto Warburg, in which cancer cells uptake glucose and use glycolysis to ferment it, and excrete the resulting lactate at very high rates, even if they have access to normal oxygen levels [2]. Despite glycolytic fermentation generating much less ATP than aerobic respiration, the Warburg effect does allow more glucose to be used for production of biosynthetic precursors, such as nucleic and amino acids [13]. These compounds are also needed for rapid cancer cell growth, and often are the limiting factor in such growth instead of ATP [13]. In addition, several other pathways, including glutaminolysis [73] and one-carbon metabolism [3], have been intensively studied in recent years, and have similarly been shown to contribute significantly to tumorigenesis.

Yet there are still many details of cancer metabolism that remain to be explored. Most importantly, there are many additional metabolic pathways that have yet to be experimentally studied for their impact on tumorigenesis. Partly, this is because the most common experimental method for studying

experimental is 13C flux analysis [74]. This method uses nutrients, such as glucose and glutamine, which have been labeled at specific positions with 13C atoms in place of 12C. By feeding such nutrients to cells in cell culture, the labeled nutrients are broken down and 13C atoms dispersed so that they label other metabolic compounds. These labeling patterns can then be measured through mass spectrometry, and metabolic flux can be computationally inferred from them. But although 13C flux analysis can give very clear results for some pathways, it has not yet been scaled up to a level that can measure metabolism across an entire cell [75].

We attempt to address this problem, by using a computational method previously developed in our lab called FALCON [65]. FALCON is part of a general class of computational methods called constraint-based modeling (CBM) [38], which model the interactions in a metabolic network as a matrix, and model constraints on metabolic flux, such as the availability of nutrients to a cell, using a series of linear inequalities. Using these concepts, FALCON then takes as input metabolic gene expression measurements from cells, maps the expression onto the metabolic network, and maximizes the linear correlation between expression and predicted metabolic flux. FALCON thus relies on the assumption that high metabolic enzyme expression is correlated with high flux, and vice versa, as described further in Methods.

We thus applied FALCON to study metabolic flux in cancer cells vs. normal controls, using data from The Cancer Genome Atlas (TCGA) [76]. Our goal was to determine which metabolic pathways were most frequently al-

tered in cancer cells, and the pattern of such alterations across the metabolic network.

3.3 Methods

We downloaded manifest files for FPKM files of mRNA expression on 17 major tissues from the GDC data portal at https://portal.gdc.cancer.gov. We chose tissues which had a minimum of 10 matched tumor and normal control samples, in order to carry out pairwise statistical tests later on. The 17 tissues are listed in 3.1, along with the corresponding number of samples for both cancer tissues and normal controls. For our analysis, we used only samples that were marked as either Primary Tumor Sample or Solid Tumor Control in our data. This excludes any samples that were marked as Metastatic.

We then mapped expression from our downloaded FPKM files onto the latest version of the human metabolic reconstruction, Recon3.02 [77]. This reconstruction models 7440 metabolic reactions that have been found in human cells. We also applied constraints to Recon2 in order to model limitations on the nutrients available to human cells. We were unable to find a source for the composition of the extracellular medium that would be located in each tissue. Therefore, we developed our own list of constraints that is intended to model the nutrients that would be available to cells grown in cell culture medium. For 43 nutrients which are listed in 3.2, chiefly amino acids,

Dataset	Number of Matched Tumor and Control Samples
BLCA	19
BRCA	113
CHOL	17
COAD	41
ESCA	11
HNSC	44
KICH	49
KIRC	72
KIRP	32
LIHC	50
LUAD	59
LUSC	49
PRAD	52
READ	11
STAD	32
THCA	58
UCEC	23

Table 3.1: TCGA datasets used in this study.

vitamins and ions, we allowed unlimited uptake or excretion into our model. For all remaining metabolites, we did not allow uptake from the extracellular medium, as we were uncertain if they were present at significant levels, but allowed possible excretion if internal reactions in our model produced them.

Metabolite Name	Metabolite Abbreviation in Recon2
glycine	gly[e]
L-argininium	arg_L[e]
L-aspartate	asp_L[e]
L-asparagine	asn_L[e]
L-cysteine	cys_L[e]
L-glutamate	glu_L[e]
L-glutamine	gln_L[e]
L-histidine	his_L[e]
trans-4-hydroxy-L-proline	4hpro_LT[e]
L-isoleucine	ile_L[e]
L-leucine	leu_L[e]
L-lysine	lys_L[e]
L-methionine	met_L[e]
L-phenylalanine	phe_L[e]
L-proline	pro_L[e]
L-serine	ser_L[e]
L-threonine	thr_L[e]
L-tryptophan	trp_L[e]
L-tyrosine	tyr_L[e]
L-valine	val_L[e]
biotin	btn[e]
choline	chol[e]
(R)-pantothenate	pnto_R[e]
folate	fol[e]
nicotinamide	ncam[e]
benzoate	bz[e]
pyridoxine	pydxn[e]
riboflavin	ribflv[e]
thiamine	thm[e]
myo-inositol	inost[e]
calcium(2+)	ca2[e]
sulfate	so4[e]
potassium	k[e]
chloride	cl[e]
sodium	na1[e]
bicarbonate	hco3[e]
hydrogen phosphate	pi[e]
reduced glutathione	gthrd[e]
oxygen	o2[e]
carbon dioxide	co2[e]

water	h2o[e]
proton	h[e]

Table 3.2: Recon2 metabolites that were allowed to be uptaken by the model in this study.

3.3.1 Description of constraints-based modeling and FALCON

Metabolic networks can be considered as bipartite networks, in which each reaction is linked to the metabolites that it consumes and/or produces [38]. CBM methods then model a metabolic network as a stoichiometric matrix S, in which each row represents a metabolite, each column a reaction, and at each row-column intersection is a coefficient representing how many molecules of a metabolite are involved in each reaction. Using this matrix, two major physicochemical constraints may be imposed. First, at steady-state, the concentrations of metabolites in a cell are neither rising or falling. Therefore, given a vector of fluxes defined as v, this constraint may be written as the equation $S \cdot v = 0$. Second, all metabolic reactions have a maximum rate, given that a cell can produce only a finite amount of enzymes, and enzymes' efficiency is limited. This constraint may be written as $v <= v_{max}$, where v_{max} is the maximum rate of a metabolic reaction.

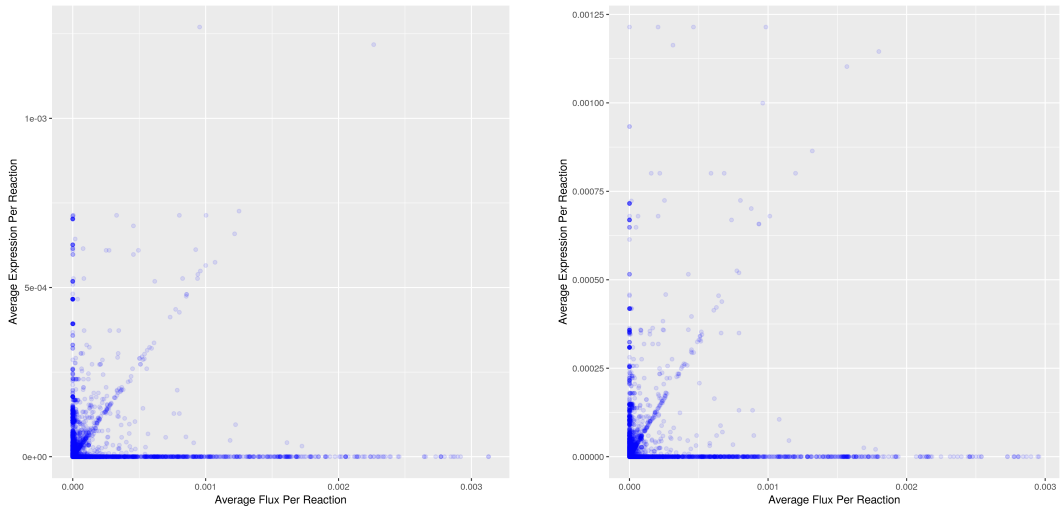
In the case of some reactions, v_{max} can be inferred based on maximum levels of experimentally measured flux rates. This is especially important for reactions which model the uptake of nutrients such as oxygen or glucose, which are often the limiting factor that constrains cell growth rates. For other reactions, v_{max} is set to an arbitrary large value, usually $+ - 1000$. This is done so that flux through these reactions will not become the limiting factor in cell growth rates.

Traditionally, metabolic network flux distributions are inferred using flux balance analysis with a biomass objective. A biomass objective represents the proper ratios of amino acids, nucleotides, and so forth necessary to create biomass, and a flux distribution that maximizes it could be said to maximize cell growth rate and therefore fitness. However, biomass optimization is generally not considered the be the most accurate way to model human cell metabolism [49]. Each tissue in humans is known to carry out distinct functions for which its metabolism is tailored, such as ion exchange in kidney cells, neuromodulator production in nerve cells, etc. During the process of tumorigenesis, cancer cells may modify normal tissue functions in order to support increased growth rates. However, mRNA and protein expression analyses of the TCGA have shown that tumor samples still retain considerable similarity to corresponding control tissues [78], and therefore metabolism of cancer tissues is also likely to retain tissue specificity. To capture this specificity more accurately, we therefore applied FALCON, which is a previously developed method in our lab that infers flux based on gene expression, by optimizing the correlation between flux and gene expression. Biomass rate is not optimized under FALCON, but instead, FALCON relies upon the assumption that high gene expression is generally correlated with high metabolic flux, and vice versa. The degree to which this assumption is true varies depending on the specific cell type and pathway. Previous investigations in bacteria have shown at least some cases where it is true, and on this basis, we have used FALCON to analyze human TCGA data as well [79] [80].

3.4 Results

3.4.1 Expression-Flux Correlations in Tumor and Control Samples

For each of the 17 tissues we examined, we first wanted to check the correlation of measured expression values with predicted flux in our FALCON simulations. FALCON is designed to maximize linear correlation of predicted flux values to expression. However, due to constraints imposed by the topology of the metabolic network, we generally find that this correlation is poor. For example, among all control samples taken from breast tissue in the TCGA dataset, we calculated the average expression and predicted flux of all 7440 reactions in the Recon2 model. We then calculated the Spearman's correlation between average expression and flux, for all reactions with either nonzero average expression or flux. There are 5518 such reactions, with a Spearman's R^2 of just .0016 (Figure 3.1a). Similar results hold for breast tumor samples (Figure 3.1b). However, among breast control samples, 5563 reactions in the model have either zero average expression or flux. These reactions are located at the bottom and left edges of Figure 3.1a. The presence of these two conflicting groups causes low expression-flux correlation, if they are included in calculating correlation. If they are excluded, then among the 1877 remaining reactions, flux-expression correlation rises to an R^2 of .244. In other words, although flux-expression correlation is globally very low across the metabolic network, there is significant flux-expression



(a) Average expression and flux of all reactions across breast control samples.

(b) Average expression and flux of all reactions across breast tumor samples.

Figure 3.1: Expression vs. flux correlation for breast samples.

correlation among the subset of the network that consistently has nonzero metabolic expression and flux.

Furthermore, we observed an interesting pattern across all TCGA datasets when we compared flux-expression correlations in tumor samples vs. normal samples (Figure 3.2). These correlations are significantly stronger in tumor samples than normal samples, across almost all datasets. When considering reactions with nonzero flux OR nonzero expression, the increase is only .0017 on average, although it applies to all datasets besides cholangiocarcinoma. However, for reactions with nonzero flux AND nonzero expression, the average increase is .083, and all datasets show such an increase except for prostate. Furthermore, the total number of reactions with nonzero expression OR nonzero flux is on average 529 reactions greater in control vs. tumor

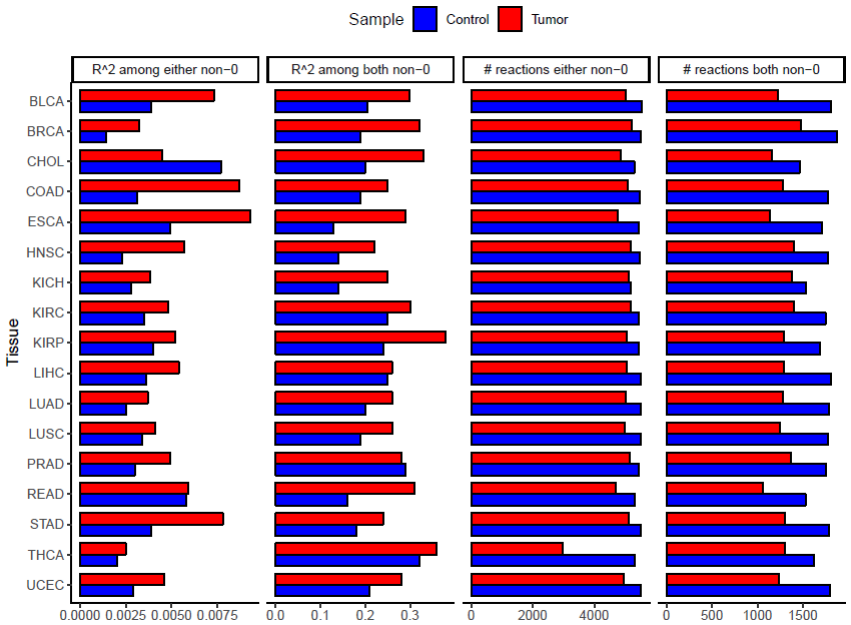


Figure 3.2: Expression-flux correlations and numbers of nonzero reactions, across 17 TCGA tissues.

samples. For nonzero expression AND flux, the difference is on average 434 more reactions in control samples.

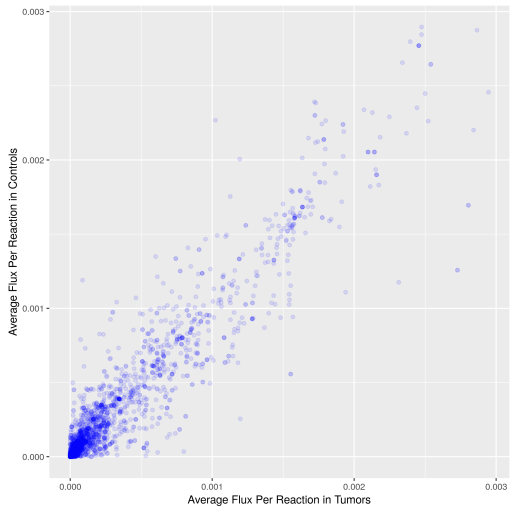
Our interpretation of these results is that tumor samples exhibit a more streamlined metabolism, with less post-transcriptional regulation than control samples. Greater expression-flux correlation indicates that regulation after transcription, such as differences in translation rate or post-translational modifications, has less impact in weakening the correlation between expression flux. This may occur because tumor tissues have large numbers of mutations throughout their genome [81], some of which may alter post-transcriptional regulation of metabolism. Furthermore, if fewer reactions have nonzero expression and flux in tumor samples, this indicates that tumor

samples have switched off some reactions that are unnecessary for proliferation. This makes sense, considering that normal tissues are specialized to carry out distinct metabolic functions, whereas cancers undergo strong positive evolution for the sole purpose of proliferating quickly, and therefore may converge on a common metabolic phenotype [82].

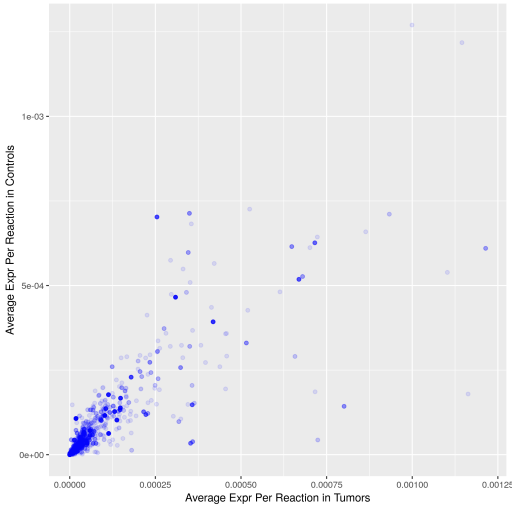
3.4.2 Flux-Flux and Expression-Expression Correlations in Tumor and Control Samples

We furthermore wanted to examine how similar flux and expression may be in normal and tumor samples of the same tissue. As an example, we plotted the average flux of each reaction in breast control and tumor samples, and calculated the Spearman correlation between the two measurements (Figure 3.3). We did the same procedure for expression measurements as well. Both analyses showed high correlations of $R^2 = .89$ for flux and .98 for expression. Furthermore, this pattern applies to all of our examined datasets, with high correlations for both flux-flux and expression-expression comparisons, and the latter being stronger. This implies that although overall metabolic expression differences between tumor and control samples may be small, they result in somewhat greater differences in flux later on. As will be seen, we also find a number of differential flux and expression subsystems in each tissue.

We also compared the flux-flux and expression-expression correlations be-



(a) Average flux of all reactions in breast control samples vs. tumor samples.



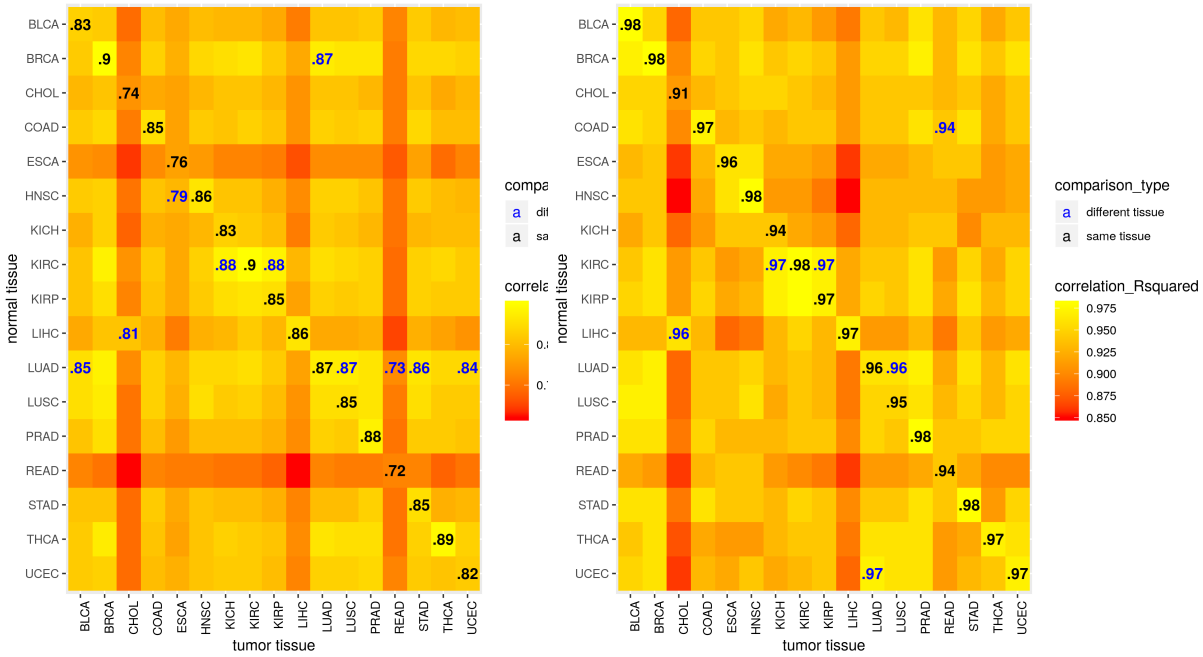
(b) Average expression of all reactions in breast control samples vs. tumor samples.

Figure 3.3: Flux vs. flux and expression vs. expression correlations for breast samples.

tween tumor samples in one tissue, versus control samples in another tissue (Figure 3.4). To our surprise, we find that in most cases, tumor samples do not show the highest flux or expression correlation with control samples in the same tissue, but instead with control samples in some other tissue. For example, tumor cholangiocarcinoma samples have a flux-flux correlation of .76 with bile duct control samples, but a correlation of .80 with liver control samples in the LIHC dataset. Although the difference is not large, this unexpected result implies that in the course of tumorigenesis, tumor samples alter their metabolic expression and flux, in such a way that it actually results in higher similarity to a different control tissue than the original tissue. Furthermore, there are a few cases where the two tissues are closely related, such as LUSC (lung squamous cell carcinoma) and LUAD (lung adenocarcinoma), but the majority of connections are between unrelated tissues.

3.4.3 Differential Metabolic Flux and Expression Subsystems

For each of the 17 tissues we examined, we used the Wilcoxon signed-rank test to determine reactions with significant differential expression or flux between tumor samples vs. normal controls. We then determined metabolic subsystems that were enriched in such reactions, by using gene set enrichment analysis (GSEA) [83]. We also repeated this analysis with expression data instead of flux data for each sample, to calculate differential expression for



(a) Correlations of average flux in control samples vs. tumor samples.

(b) Correlations of average expression in control samples vs. tumor samples.

Figure 3.4: Correlations of tumor flux vs. control flux and tumor expression vs. control expression across all 17 TCGA tissues.

all reactions and pathways in each tissue. For both of these analyses, we considered a subsystem to be enriched in a tisue if its GSEA p-value was less than .05.

Overall, thirty metabolic subsystems were enriched in reactions with higher flux in tumors in at least 1 tissue. Our results highlight several pathways with high differential flux that are already known to be important in tumorigenesis. The most common differential flux subsystem with higher flux in tumors, in eight out of seventeen tissues, is methionine and cysteine metabolism, which involves reactions for biosynthesis of these two closely related amino acids. Both of them are required inputs to folate and one-carbon metabolism, which are further linked to DNA methylation and nucleotide synthesis pathways. A recent paper has shown that methionine restriction has an anti-tumorigenic effect [84], and therefore it is not surprising that higher biosynthetic flux is observed in tumors. Furthermore, the next most common subsystem with higher flux in tumors is tyrosine metabolism, which also involves biosynthesis of a key essential amino acid. Fatty acid synthesis has higher tumor flux, because it is necessary for increased cell membrane size, which expands greatly in area to accompany cell proliferation in tumors. Sphingolipid metabolism also had higher differential flux in several tissues, which may be related to the properties of sphingolipids as key signaling metabolites, with either an pro- or anti-tumorigenic effect in many tumors [85]. Production of the sphingolipid ceramide is associated with increased apoptosis, autophagy, and cell death, while the related sphin-

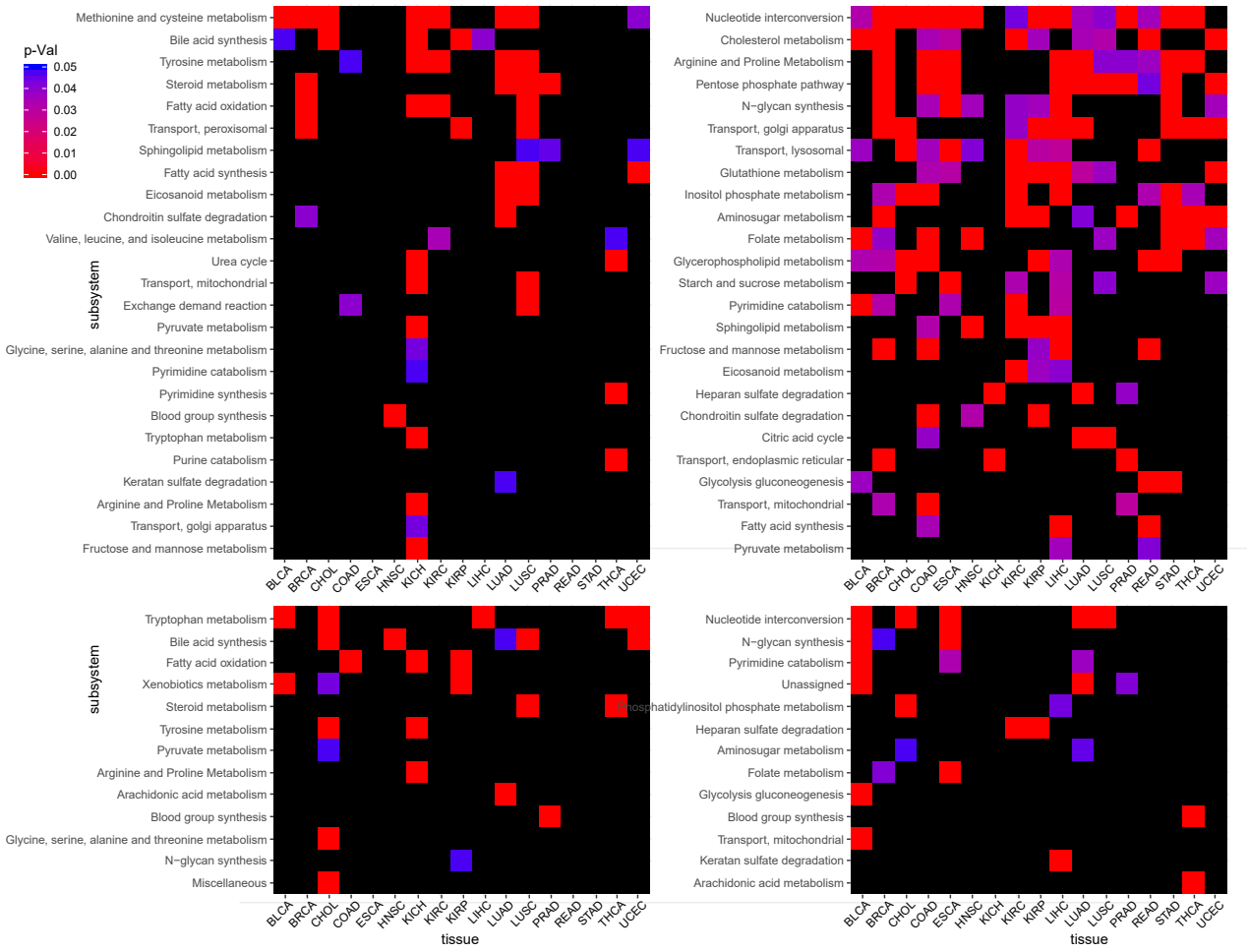


Figure 3.5: Top-left: GSEA positive differential flux subsystems. Top-right: GSEA negative differential flux subsystems. Bottom-left: GSEA positive differential expression subsystems. Bottom-right: GSEA negative differential expression subsystems.

golipid sphingosine-1-phosphate activates oncogenic signaling receptors [85]. Finally, bile acid metabolism shows higher flux metabolism in tumors, which is highly interesting given a recent paper on how the gut microbiome may modify primary bile acids that are initially produced by the liver into secondary bile acids. These secondary bile acids have been shown to promote liver tumorigenesis by altering the immune response in liver tissues [86].

Twenty-six subsystems were enriched in reactions with lower tumor flux, in at least one tissue. Among the most common subsystems among all tissues are nucleotide interconversion and the pentose phosphate pathway. Both of these are involved in the supply of nucleotides for DNA synthesis, so it is surprising that they exhibit lower flux in tumors. However, one possibility is that tumors instead obtain most of these compounds by scavenging from the extracellular environment, which has been previously been shown to be a major contributor to cancer metabolism [87]. Similarly, although cholesterol metabolism flux is lower in most tumors compared to controls, this may be explained by higher tumor scavenging of cholesterol. Cholesterol is a major component of cell membranes, and is also important in cancer signaling related either to lipid rafts or mTORC1 [88]. On the other hand, N-glycan synthesis, aminosugar metabolism, and inositol phosphate metabolism all exhibit lower flux, and in this case may be expected. All three subsystems are related to the synthesis of glycoproteins which frequently serve as signaling markers on the cell surface, and the composition of glycoproteins has been shown to be altered in cancers [89]. Finally, starch and sucrose metabolism,

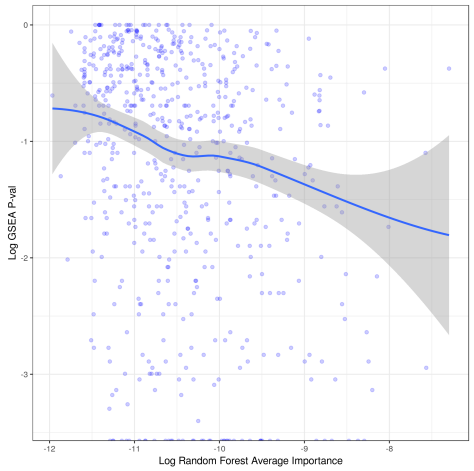
fructose and mannose metabolism, and pyrimidine metabolism are reduced as well, further suggesting that cancers may reduce their dependence on these nutrient sources, in favor of others.

Furthermore, when we look at differential expression subsystems in our dataset, we saw first there were considerably fewer such subsystems, with only seventeen with lower tumor expression in at least one tissue, and fourteen with higher. Nevertheless, a few of these overlap with the differential flux subsystems previously described, among them sphingolipid metabolism and bile acid synthesis with higher flux and expression in tumors, and nucleotide interconversion, pyrimidine catabolism, N-glycan synthesis and aminosugar metabolism with lower flux and expression. These results give us added confidence that the subsystems in question are indeed upregulated in cancer metabolism, and play a significant role in tumorigenesis. On the other hand, the remaining differential flux subsystems, which do not have differential expression, indicate that FALCON may be able to independently predict significant flux differences, and thus give additional information on metabolism, outside of expression measurements. But furthermore, there are also several subsystems with strong differential expression, but without corresponding differential flux. Among those with higher tumor expression are tryptophan metabolism, which involves an essential amino acid, and steroid metabolism, which may be important in cancer signaling networks. Folate metabolism has lower tumor expression, though it has been shown to be essential in maintaining tumor nucleotide synthesis and methylation. These subsystems

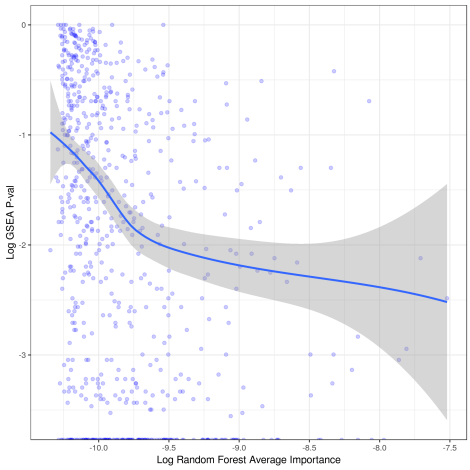
probably do correspond to real cases of differential flux in cancer tissues, but they are not predicted by FALCON. One reason may be that mRNA expression is known to not be perfectly correlated with flux [80], and FALCON may therefore perform better if given access to proteomics data, which is believed to be more relevant for flux prediction.

To further test whether we obtain significant flux and expression differences between tumor and control samples, we also used random forests to extract significant reactions. We used the package randomForests in R, with the parameters ntrees=501, importance=TRUE, proximities=501, to distinguish normal from tumor samples in all 17 tissues. From the out-of-bag error of the trees, we found that classification accuracy was very high in all tissues, using both flux and expression data, with an average of 96% accuracy. However, if we rank all metabolic reactions by their importance in classifications, according to mean decrease in accuracy when the reaction is excluded, we find that there is a very smooth decline in importance. We therefore decided to rank metabolic subsystems based on their average importance. We calculated the average importance of reactions within each subsystem, and used this to rank each subsystem.

Importantly, there is a weak, but significant negative correlation, between the average importance of a subsystem according to random forests, and its p-value in GSEA. Specifically, for each subsystem in Recon2, and each tissue for which control and tumor samples, we obtained both the average importance of its reaction fluxes in random forests, as well as its p-value



(a) Scatterplot of average log GSEA p-val for expression values of all reactions, versus average log random forest importance.

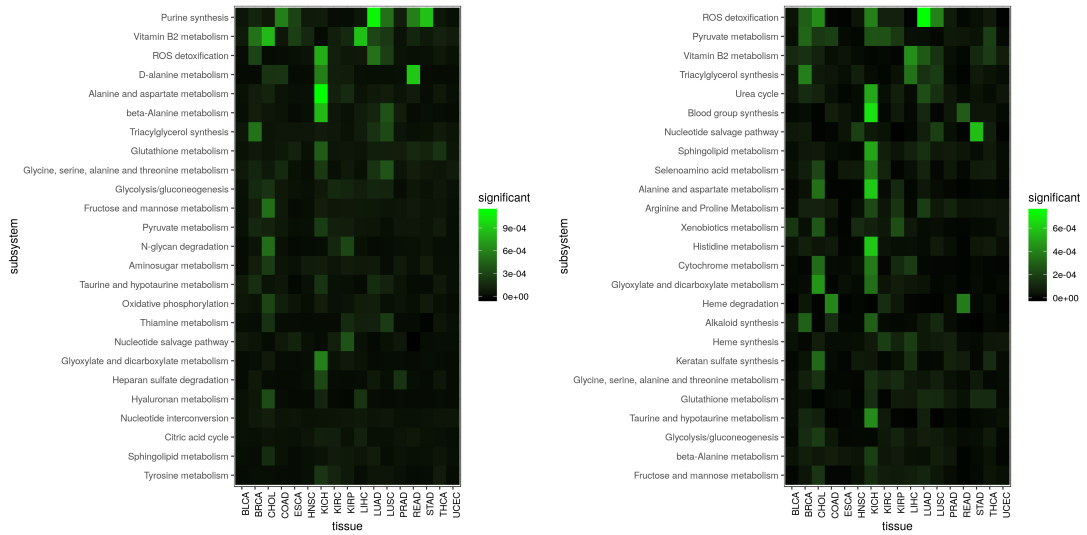


(b) Scatterplot of average log GSEA p-val for fluxes of all reactions, versus average log random forest importance. Lowess regression line and confidence intervals are also shown.

Figure 3.6: Scatterplots of average log GSEA p-val versus average log random forest importance, for either expression values or fluxes of all reactions.

in GSEA. We computed the correlation of these two measures, as well as for expression measurements. The results are plotted in Figure 3.6, and show that the Spearman's R^2 values are .07 for flux correlation and .15 for expression measurements, which although low have corresponding p-values of 3.80e-12 and 9.23e-28.

The most important subsystems for flux differentiation by this metric include a few overlaps with the GSEA results. These include aminosugar metabolism, N-glycan degradation, and fructose and mannose metabolism. However, most of the highest-ranked flux subsystems are different from the GSEA results. They include some that are likely important for tumorigenesis,



(a) Heatmap of top differential flux subsystems according to random forest across all tissues.

(b) Heatmap of top differential expression subsystems according to random forest across all tissues.

Figure 3.7: Heatmaps of top differential subsystems by random forests.

such as ROS detoxification and glutathione metabolism, which help deal with greatly elevated ROS levels in tumors [1], and vitamin B2 metabolism, which is a critical cofactor for several key metabolic enzymes [91]. Notably, ROS detoxification and B2 metabolism are also important when ranking based on average random forest importance using expression data.

Finally, we also wanted to test for differential flux in transport of extracellular metabolites in tumor vs. normal control tissues. We wanted to focus specifically on these metabolites, because uptake or release of specific metabolites can sometimes be measured in vivo in cancer patients, and may serve as a crucial biomarker. An important example of this is the ingestion of labeled glucose tablets, whose uptake by cancer cells can be monitored by

FDG-PET to measure the rate of glucose uptake in general, a key measure of tumor malignancy [13]. Therefore, we examined the top 30 metabolites that most commonly had differential flux or expression, according to the Wilcoxon signed-rank test, across all 17 tissues.

For metabolites whose uptake into cancer cells was predicted to be greater, we observed that among the most common predictions were two biosynthetic compounds, AICAR, an essential intermediary in purine synthesis, and phosphatidylglycerol, a component in the head groups of many membraine lipids. Also common were thromboxane, an eicosanoid, and putrescine, a compound related to spermidine, which may play roles in cancer signaling networks.

Among metabolites whose uptake into cancer cells was predicted to be less, there were two adenine-related compounds, ADP-ribose and cyclic-AMP. cyclic-AMP also plays a major role in many cell signaling pathways. Additionally, arabinose, octanoate, and two compounds related to linolenic also had less flux, suggesting that cancer cells utilize these potential nutrients less.

3.5 Discussion

Altered metabolism is known to be one of the major characteristics of cancer tissues compared to non-cancerous tissues. Despite numerous small-scale experimental studies, changes in metabolic flux in cancer have never been studied experimentally on a genome-wide scale before. Therefore, we applied a

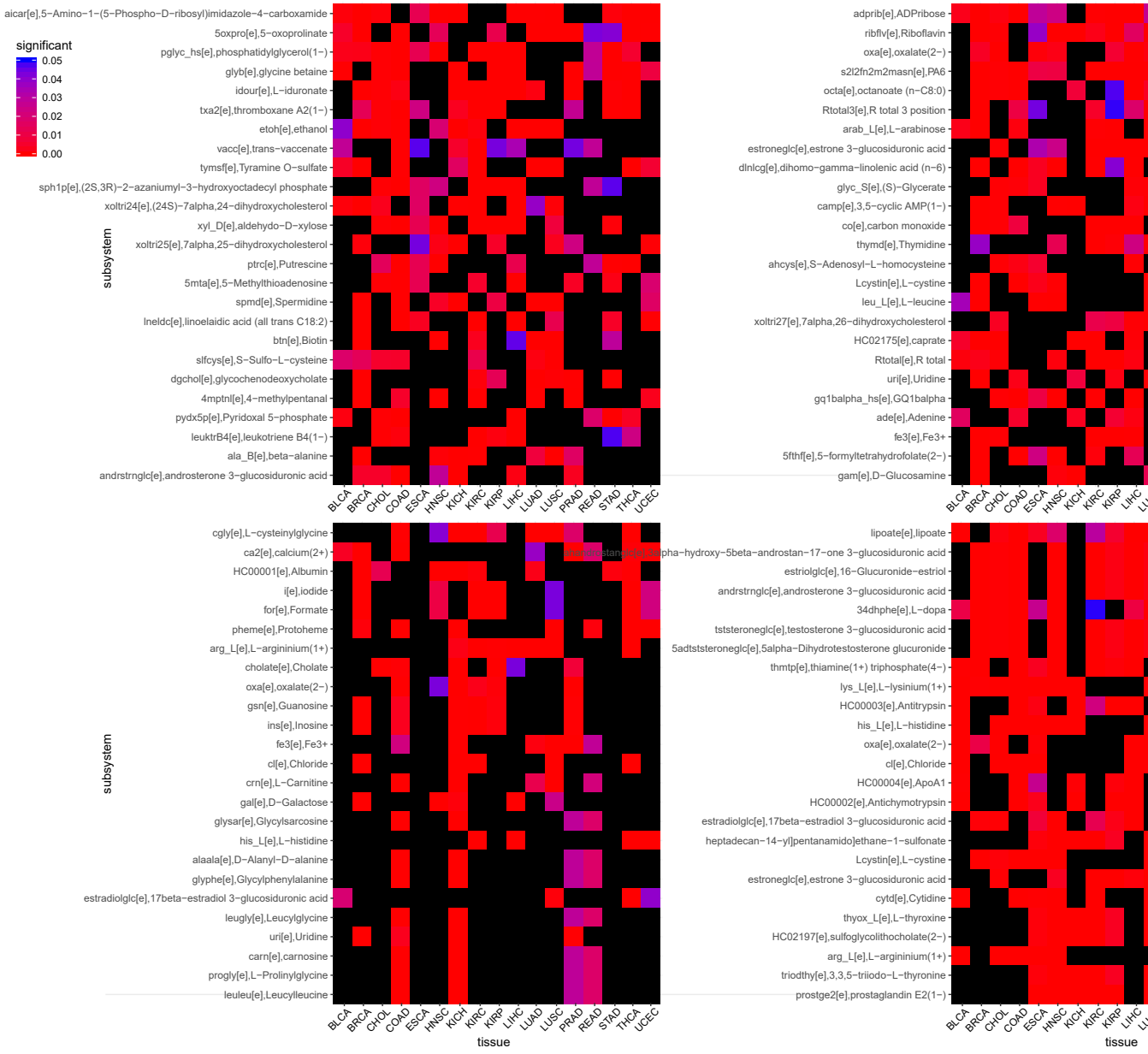


Figure 3.8: Top-left: Positive differential flux transport reactions. Top-right: Negative differential flux transport reactions. Bottom-left: Positive differential expression reactions. Bottom-right: Negative differential expression reactions.

computational method called FALCON, developed in our lab previously [65], with the previously published human metabolic model Recon2 [77], to infer flux differences from RNA-seq expression data in the TCGA project.

We first identified several interesting relationships between expression and inferred flux in the TCGA dataset. Although the correlation is in general weak, it is much stronger among the subset of reactions in Recon2 that always have nonzero expression and flux. Furthermore, this correlation is considerably stronger in tumor tissues than their normal control counterparts, and tumor tissues also display a significant smaller number of reactions with nonzero expression and flux. These results imply that tumor tissues have a more streamlined metabolism, with fewer layers of regulation between expression and flux, especially among the most active subset of metabolic reactions.

Our work also helps to identify a large number of candidate metabolic pathways that may be altered in cancer tissues. Although further experimental work must be done to validate whether any of these represent meaningful differences, many of them have previously been shown to involved in tumorigenesis. Furthermore, we also used two different methods for determining differential flux subsystems, GSEA and random forests, and also determined differential expression subsystems. We can use results from all these methods together to focus on subsystems that are important by all measures, and therefore more likely to reflect true biological importance.

Finally, we also tested for significant metabolites whose exchange fluxes are different between normal and tumor tissues. These results also suggest

numerous candidate metabolites whose uptake or excretion may be different between cancer and normal tissues.

In total, our work attempts to determine significant metabolic flux differences between cancer and tumor tissues, based on flux inference by the FALCON algorithm. We observed several major pathways that have both differential flux and expression, as well as several metabolites whose uptake and release are predicted to be different, and are linked with the differential pathways. We expect that future studies may further improve the accuracy of metabolic inference in cancer.

Chapter 4

Expression-Based Inference of Human Microbiome Metabolic Flux Patterns in Health and Disease

4.1 Abstract

Metagenomic sequencing has revealed that the composition of the gut microbiome is linked to several major metabolic diseases, including obesity, type 2 diabetes (T2D), and inflammatory bowel disease (IBD). However, the exact mechanistic link between the gut microbiome and human host phenotypes is unclear. Here we used constraint-based modeling of the gut microbiome, using a gene-expression based algorithm called FALCON, to simulate metabolic flux differences in the microbiome of controls vs. metabolic disease patients. We discovered that several major pathways, previously shown to be important in human host metabolism, have significantly different flux between

the two groups. We also modeled metabolic cooperation and competition between pairs of species in the microbiome, and use this to determine the compositional stability of the microbiome. We find that that the microbiome is generally unstable across controls as well as metabolic microbiomes, and metabolic disease microbiomes even more unstable than controls.

4.2 Introduction

Metabolic flux is one of the most important phenotypes that can be measured for a cell [74]. It determines what types of nutrients the cell will uptake, and how it will use them to harvest energy and create the building blocks of a cell such as nucleic and amino acids [92]. This determines the growth rate of a cell and ultimately its fitness in an environment with a limited supply of nutrients. A large part of the cell's regulatory machinery is therefore devoted to regulating metabolism [93], and changes in metabolism play an important role in the development of many diseases, such as cancer [94], diabetes [95], and heart disease [96].

In recent years, another area where metabolism has been shown to be important is the human gut microbiome, which has seen great interest for its impact on nearly every aspect of health of its host [97]. It has been shown that the gut contains a variety of microorganisms, almost all bacterial, which in total have roughly 10 times the cell count of the human host, and 100 times as many genes [24]. The gut microbiome carries out multiple

important functions for its human host, including digestion of fibers that can contribute up to 10% of the body's energy needs, synthesis of essential vitamins, and competition with pathogens that could infect humans [98]. Not surprisingly, then, disruptions in the microbiome's composition have been linked to numerous diseases, including gut diseases such as colon cancer and Crohn's disease [99], as well as many that may seem unrelated to it, such as atherosclerosis [100]. Typically, these links are inferred by a technique similar to GWAS, by measuring species composition based on sequencing of the 16s rRNA gene, and comparing this composition between healthy and diseased individuals [101].

However, even though microbiome composition may be clearly associated with disease, it is still unclear whether there is any association with changes in microbiome metabolic flux. Given the crucial role of gut-derived metabolites, such as SCFAs [102], in the human body, it seems reasonable to assume that changes in gut microbiome metabolic flux play an important role in the development of certain diseases. One of the most striking findings from the recently completed Human Microbiome Project is that healthy subjects may have very different gut microbiome compositions, but functionally had almost the same metagenomic coverage of metabolic pathways [103]. However, metabolic flux also remains one of the most difficult phenotypes to experimentally measure. The current state-of-the-art is to use 13C isotopes to label metabolites in the cell, followed by mass spectrometry to measure the concentration of labeled metabolites. The overall labeling patterns can be used to

computationally infer flux [104]. These methods suffer from two major limitations. First, mass spectrometry cannot accurately measure low-abundance labeled metabolites, which may be crucial for accurate flux determination on a genomic scale [75]. Second and more importantly, mass spectrometry measurements usually involve an average across a sample of cells [75], involving many different species in the gut microbiome, and so cannot resolve the metabolic state of individual species.

To overcome these challenges, several previous groups [105] [106] have used a class of algorithms called constraint-based modeling (CBM) [38] to computationally infer metabolic flux in the microbiome. As described further in the Methods section, CBM methods are designed to take a minimal set of physical and chemical assumptions (constraints) about metabolic flux, and use these to determine a reasonable solution space of metabolic flux distributions. However, in order to deal with the presence of multiple species in the gut microbiome, these previous methods must make additional assumptions about metabolic interactions between species, also described further in Methods. We have therefore developed two innovations in our paper to more accurately model gut microbiome metabolism. First, we applied a previously developed CBM method in our lab called FALCON [65]. FALCON takes the solution space determined as above, and then determines one particular, optimal flux distribution, by optimizing the agreement between the metabolic flux of each reaction, and the enzymatic expression associated with that reaction. This approach follows the general principle that highly active

reactions require large amounts of expressed enzyme in order to function, and vice versa [79]. We believe that FALCON requires fewer assumptions than previous work, and therefore may be more suited to modeling microbiome metabolism.

Second, we also used three different types of metabolic network models for the microbiome, each of which captures different types of important interactions. As described below, we first used the single species AGORA models [107], a set of 773 metabolic reconstructions for individual gut microbiome species. The strength of these models is that they allow us to analyze major flux differences between samples, and break down which individual species are responsible for these changes. However, a drawback of the single species models is that they cannot capture any metabolic interactions between species. To capture these, we also developed a merged gut microbiome model, in which all reactions in the AGORA models are connected into a single network of 3670 reactions. Although the merged model loses all species specificity, it allows modeling of metabolic interactions among any number of microbiome species, as all reactions are included within it. Finally, between single-species and merged models, we also examined all pairwise models, in which all possible pairs of AGORA models in a sample are placed together in a common environment. This allows us to capture a subset of metabolic interactions, those between all pairs, while still retaining species specificity. As will be seen, the pairwise models are also sufficient to investigate the compositional stability of the gut microbiome.

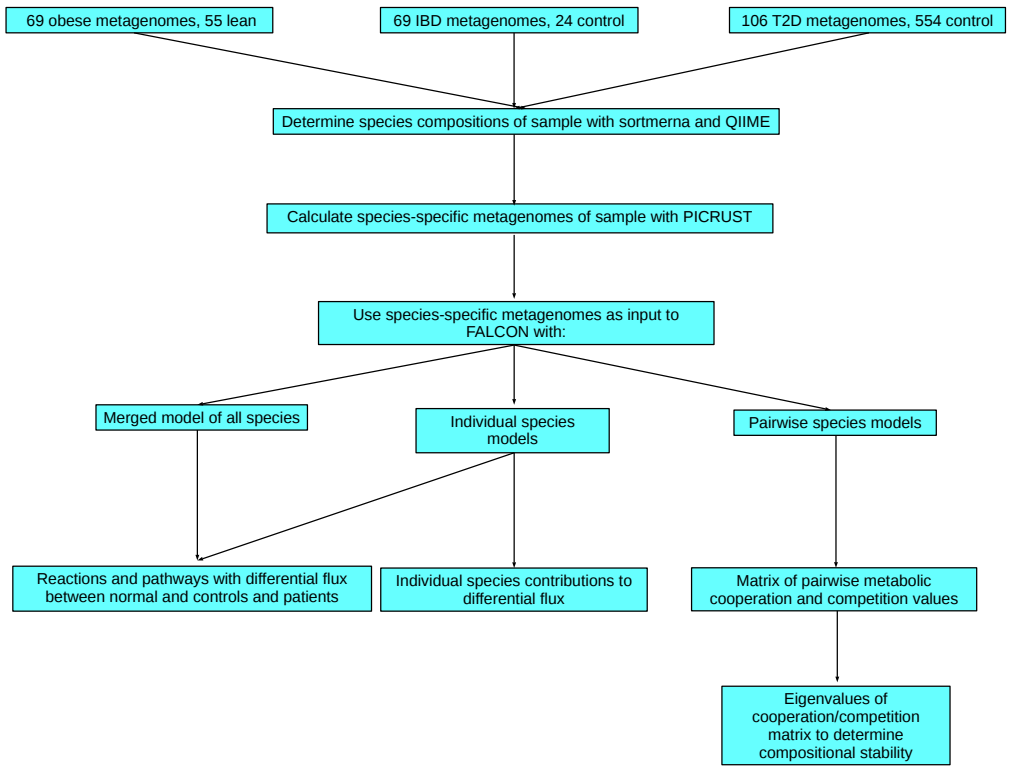


Figure 4.1: Flowchart of analyses carried out in this paper.

We applied FALCON to analyze the metagenomic expression data for microbiomes from patients of three different metabolic diseases, namely obesity [26], Type 2 diabetes [32], and IBD [33], using each of the three models we have described. We calculated the resulting flux distributions in all cases, and used the most appropriate model to determine the pathways which carry different amounts of flux between normal and obese subjects, as well as stability differences between them. A flow chart of our workflow is shown in Figure 4.1.

4.3 Methods

Metabolic networks can be considered as bipartite networks, in which each reaction is linked to the metabolites that it consumes and/or produces [38]. CBM methods thus model a metabolic network as a stoichiometric matrix S, in which each row represents a metabolite, each column a reaction, and at each row-column intersection is a coefficient representing how many molecules of a metabolite are involved in each reaction. Using this matrix, two major physicochemical constraints may be imposed. First, at steady-state, the concentrations of metabolites in a cell are neither rising or falling. Therefore, given a vector of fluxes defined as v, this constraint may be written as the equation $S \cdot v = 0$. Second, all metabolic reactions have a maximum rate, given that a cell can produce only a finite amount of enzymes, and enzymes' efficiency is limited. This constraint may be written as $v <= v_{max}$, where v_{max} is the maximum rate of a metabolic reaction.

In the case of some reactions, v_{max} can be inferred based on maximum levels of experimentally measured flux rates [92]. This is especially important for reactions which model the uptake of nutrients such as oxygen or glucose, which are often the limiting factor that constrains cell growth rates. For other reactions, v_{max} is set to an arbitrary large value, usually $+ - 1000$. This is done so that flux through these reactions will not become the limiting factor in cell growth rates [38].

Traditionally, the distribution of in a metabolic network inferred using

flux balance analysis with a biomass objective. A biomass objective is a special reaction in the network represents the proper ratios of amino acids, nucleotides, and other metabolites necessary to create biomass. A flux distribution that maximizes the rate of the biomass reaction is considered to maximize cell growth rate, and therefore fitness. However, biomass optimization may not be the most accurate way to model metabolism in the microbiome, for several reasons. It is known that the human host can alter the composition of the microbiome through mechanisms such as immune targeting, and provision of nutrients like mucins to certain species [108]. Some bacterial species may follow strategies besides biomass optimization to survive in the microbiome, such as producing toxins that reduce growth of competitors [108]. Finally, in any simulation method involving biomass optimization in the microbiome, different species would be expected to have different biomass growth rates. This implies that species which grow the fastest would be able to outcompete other species in the microbiome, and eventually dominate the gut environment. As this is not observed experimentally, previous methods that simulate gut metabolism with FBA must incorporate additional constraints to prevent unbalanced growth rates [106] [105].

To avoid these problems, we instead chose to apply FALCON, which is a previously developed method in our lab that infers flux based on gene expression, by optimizing the correlation between flux and gene expression. Details of the FALCON algorithm may be found in [65]. Biomass rate is not optimized under FALCON, and indeed, in our FALCON simulations, we have

never observed flux through the biomass reaction. However, the rationale behind FALCON is that high enzyme expression generally correlates with high flux, and vice versa [79]. Although this correlation is very weak in some cases, and has not been experimentally measured in the gut microbiome, we still believe that FALCON may provide a more principled simulation method than traditional FBA.

We furthermore applied FALCON to three different types of metabolic models for the gut microbiome, all of them based on the previously published AGORA database of gut microbiome models. The AGORA models are a set of 773 metabolic reconstructions for gut microbiome species. Importantly, they contain a set of estimated lower and upper bounds on uptake of common nutrients, such as glucose and amino acids, that are representative of a Western diet [107]. We used these limits for all simulations in our study.

Each of the three types of AGORA-derived metabolic models has their own strengths and drawbacks. First, we applied FALCON to the individual species-specific models from AGORA. This setting represents how metabolic flux would be distributed in each species if they were growing alone. Although simulations with these models do not capture any species interactions, they do show whether any particular species has significant metabolic differences between normal control and metabolic disease samples. They may also be useful to study the metabolism of individual microbiome species which can be cultured. Second, we applied FALCON to pairwise models, in which we modeled every possible pair of species, placed together in a common environ-

ment, i. e. sharing common nutrient uptake and waste product secretion, and then simulated with FALCON. As will be seen later, this is the most efficient way for us to elucidate all pairwise interactions among species, which can be used to calculate the compositional stability of the microbiome as a whole.

Finally, we merged all microbiome models in AGORA together, by combining all unique metabolic reactions across all species into one model, to form a very large merged model of 3670 reactions, representing the combined metabolic potential of the gut microbiome. The advantage of using this model is that it is able to capture how metabolic fluxes in the entire microbiome are distributed, taking into account all possible species interactions. That is, individual species in the gut may strongly interact with each other in groups of three or more species, and it is impossible to individually model all such possible groups. However, the merged model allows all reactions in all species to be linked to each other, allowing interactions across the entire microbiome to be modeled.

Furthermore, the size of the merged model depends on which species are included, and it should be noted that the size quickly saturates as more individual models are included. This is shown in the left half of Figure 4.2, which shows the number of reactions in the merged model across 1000 randomized trials, in which models are selected from the AGORA set and added in a random order. The result is important, as the AGORA set does not cover all known gut microbial species, and potentially new species added to the model may add new reactions. We do not expect this to be a major draw-

back though, as such new reactions would only comprise a small fraction of the merged model's total. Furthermore, the metagenomic samples we consider may have very different species compositions, and thus the number of metabolic models that can be imputed to each model may be very different, as shown in the right half of Figure 4.2. This figure shows the number of individual metabolic models present within each sample of our diabetes dataset. However, similar to the findings of the Human Microbiome Project [24],we would also expect that each sample should have very roughly constant coverage of metabolic pathways, as all microbiomes must carry out the same set of metabolic functions. Indeed, when we constructed the merged models for each of our diabetes samples, by merging all individual models in the sample, we found that they all contained almost the same number of reactions. We interpret this to mean that all samples do have complete coverage of all metabolic reactions in the microbiome, as expected from the result in Figure 4.2

Finally, in order to map metagenomic reads for each species to determine their abundance, we followed two different procedures among our three disease-specific datasets. For both IBD and diabetes datasets, we downloaded appropriate tables of species abundances, based on 16s reads, from their supplementary data. For our obesity dataset, only metagenomic samples were obtained, and such a table was not available, so we used the following pipeline to extract species abundances. First, we first filtered for 16S reads in each metagenomic sample of the obesity dataset using sortmerna [109].

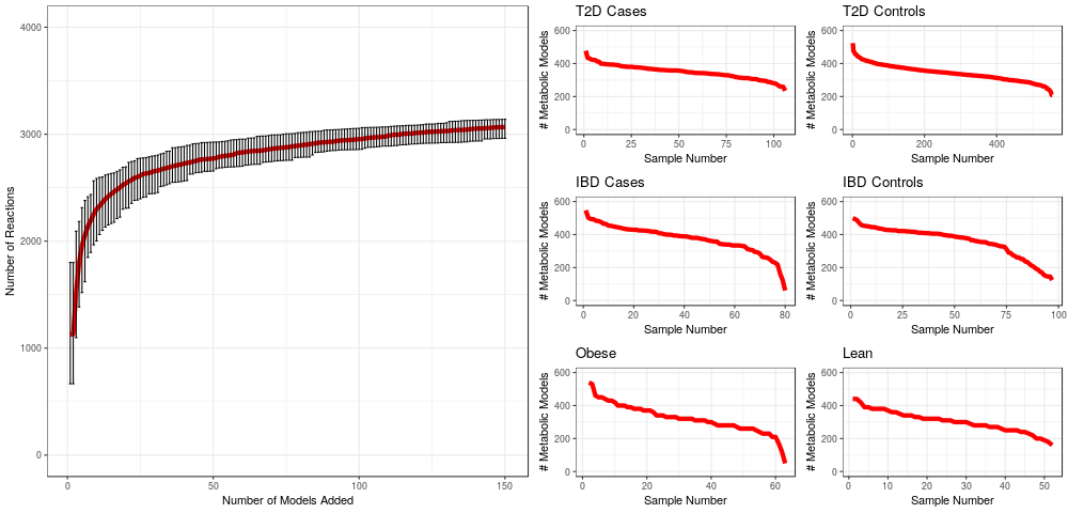


Figure 4.2: Left: For 1000 iterations, individual AGORA models were added to merged model in randomized order. Median, minimum, and maximum number of reactions in the merged model, after adding up to 150 individual models, are shown. Right: Number of individual metabolic models that were found to be present in each sample, from the three microbiome studies that were examined in this paper.

We ran this program with the following parameters: using the SILVA 16s database [110], and we requiring a 90% match threshold of reads to a 16s dataset. We then used the QIIME2 [111] command pick_closed_reference.py to map the filtered 16S reads, output by sortmerna, against the Greengenes database taxonomy [112]. This results in species abundances based on 16s reads from the obesity dataset.

With the 16S abundances of each species in each study, we then used PICRUST [113] to infer abundances of each species-specific metagenome, with the default parameters as given in the paper. These metagenomes were used in FALCON to simulate the metabolism of single gut microbiome

species. For paired-model simulations, we used the inferred metagenomes of both modeled species. Finally, for merged model simulations, we added together the abundances of all inferred metagenomes. That is, for a reaction i in the merged model, we determined all individual models that contained reaction i, along with reaction i's inferred metagenomic abundance in each individual model. We added together the inferred abundances of reaction i, and used that as the abundance of reaction i in the merged model. We used this method, as opposed to simply mapping metagenomic reads to the merged model, because all species in the gut microbiome would contribute to metagenomic reads, including some not included in the AGORA models. However, the merged model was initially created from individual AGORA models, and we only wish to include the metagenomic abundances of those individual AGORA models.

4.3.1 Calculation of Taxonomic Distances Between Individual-Species Models

We downloaded version 13.5 of the Greengenes taxonomy [114]. We matched the names of individual AGORA species models to leaf nodes of the taxonomy, and then calculated the taxonomic distance between every pair of such leaf nodes. We used these pairwise taxonomic distances to examine the correlation of metabolic cooperation/competition with taxonomic distance (see Pairwise Models section of Results).

4.4 Results

4.4.1 Merged and Individual-Species Models

We performed simulations on metagenomic data from several studies that examined three different metabolic diseases: obesity [26], T2D [32], and IBD [33]. Using the pipeline described in Methods, we extracted species-specific metagenomes with PICRUST. We then mapped expression from each of these three datasets onto two different types of metabolic models, first the set of all individual species models present in a sample, and then the merged model of all individual species models. This results in a total of six different cases, and we performed FALCON simulations on each of them. To determine which subsystems have differential flux between normal control and patient samples, we first analyzed individual reactions. We first ranked all reactions with differential flux, based on a Wilcoxon rank-sum test for differential flux between control and diseased groups, and determined all those reactions with a Wilcoxon p-value $< .05$ for differential flux. For each comparison between metabolic diseases and controls, there were hundreds of such nominally significant reactions, as we did not correct for multiple testing. It should be noted, though, that the actual flux difference of flux difference between the two groups was quite small for all reactions. For example, Figure 4.3 shows the distribution of fluxes between diabetics and normal controls for a very highly differential reaction in flux, phosphoglycerate kinase, showing a small but significant decrease of .00226 flux units (Wilcoxon p-value .021) in di-

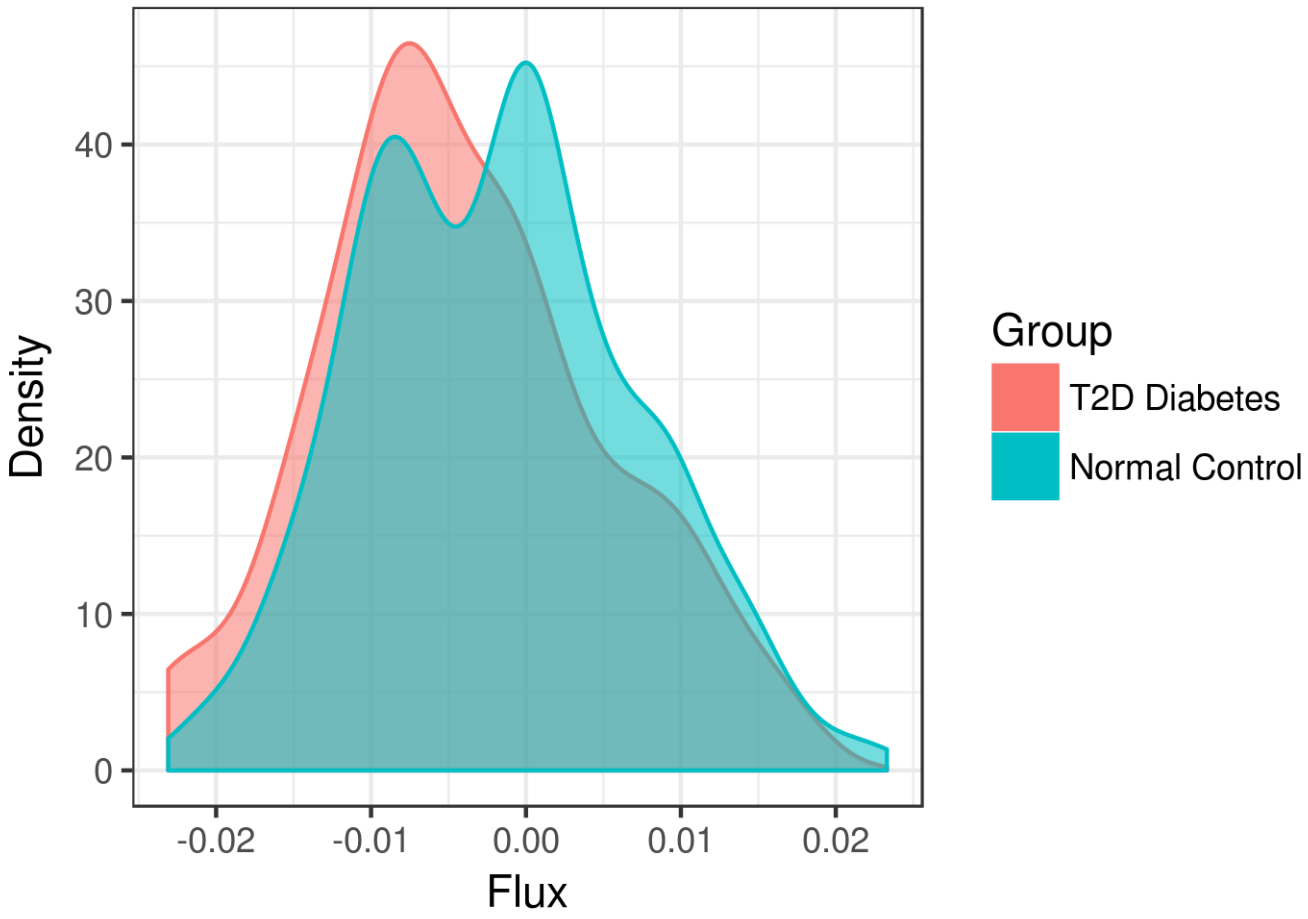


Figure 4.3: Distribution of fluxes for phosphoglycerate kinase, simulated using the merged model, among 106 T2 diabetic samples and 554 normal controls.

abetics. Finally, we calculated the enrichment of each subsystem for these differential flux reactions, based on a hypergeometric test.

The results are summarized in Figure 4.4 below for the case of the merged model simulations. Several interesting findings emerge when we compare across the three diseases. The pentose phosphate pathway, lipopolysaccha-

ride biosynthesis, and lysine, branched amino acid, and histidine metabolism are all downregulated in expression across at least one dataset. These pathways are all related to biosynthesis in some way, and thus their downregulation in patients suggests that the gut microbiota grows at a slower rate. It is also interesting to note that branched chain amino acid synthesis is downregulated in T2D patients versus normal controls, as this supports evidence that shows higher levels of such acids in diabetic patient bloodstream. Our analysis suggests that the gut microbiome may be a contributor to this difference. It is known that gut microbiome metabolism can be extensively intertwined with that of the human host. Metabolites like butyrate can be exported from the microbiome to the host, where it serves as an energy source for colonocytes [115], and conversely nutients such as mucin can be provided by the host to the microbiome [116]. On the other hand, the expression of starch and sucrose metabolism and fatty acid oxidation is upregulated in some patients, suggesting increased use of these metabolites for energy.

Finally, significant differential flux pathways were fewer than differential expression pathways. Plant polysaccharide degradation was upregulated in IBD patients, while folate metabolism and glucuronate interconversion were downregulated in obese and T2D patients respectively. Polysaccharide degradation is surprising to see as upregulated in patients, because it serves as a precursor to synthesis of SCFAs such as butyrate, which are associated with a healthy microbiome.

We were also interested in possible flux differences between the single-

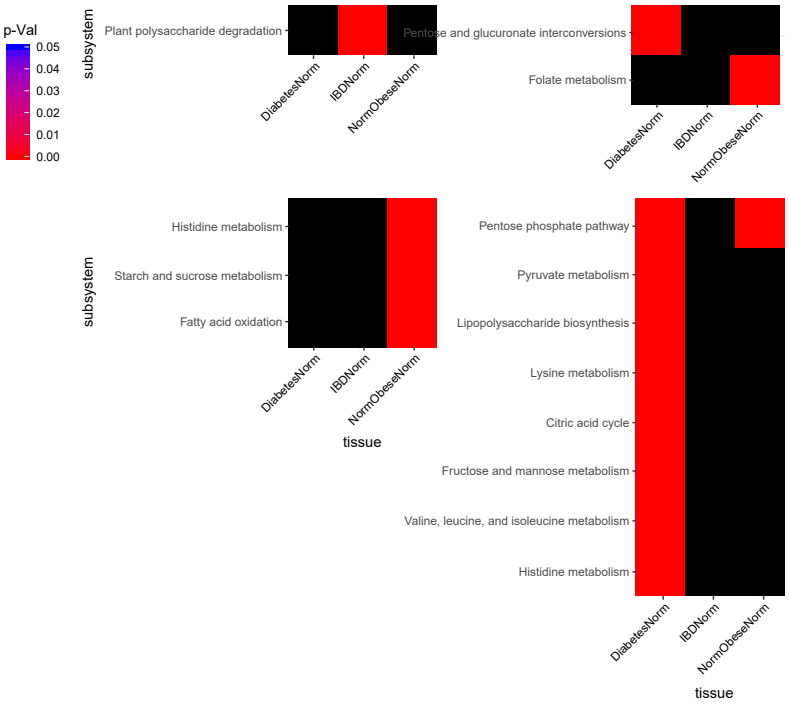


Figure 4.4: Top-left: Positive differential flux subsystems. Top-right: Negative differential flux subsystems. Bottom-left: Positive differential expression subsystems. Bottom-right: Negative differential expression subsystems.

species AGORA models, vs. the merged model. We expect that flux predictions from the two types of models would be very different, as the merged model is able to capture all possible species interactions, compared to individual species models which cannot capture any. We therefore compared the predictions for differential flux between the merged and single-species models, using each of our three datasets. In general, there was very poor correlation of individual fluxes between the two simulation cases. The merged model is expected to be more accurate in this case because it can capture more species interactions.

The poor correlation between individual species and merged models can be seen in Figure 4.5. In these plots, we considered respectively reactions that showed either a positive or negative merged-model flux difference in normal control samples vs. patients. We also determined the corresponding reactions in the single-species simulations, and the sum of either positive or negative individual species fluxes for each reaction. Then, we correlated merged-model positive differential flux with individual species positive differential flux, and similarly for negative differential flux. As can be seen, in Figure 4.5, there does not seem to be a significant correlation between the two quantities in any dataset. This indicates that differential flux in the merged model does not have a significant association with individual species differential flux.

An even more fine-grained analysis is shown in Figure 4.6, which consider only reactions that showed significant differential flux in both the merged model and single-species cases. In this case, the x-axis of the heatmap indi-

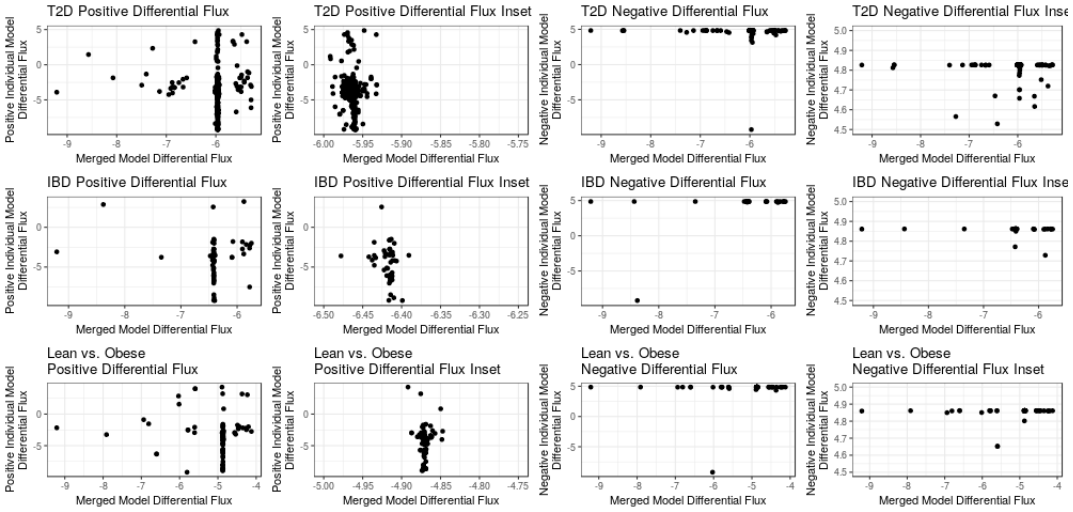


Figure 4.5: Average merged model differential flux, vs. average individual species differential flux, across samples from the three studies that were examined. Inset panels are focused on areas where data points were clustered together, in order to a get a clearer view.

cates the number of reactions in each dataset with either positive or negative merged model differential flux (Wilcoxon p-value $< .05$). On the y-axis is shown the individual species log differential flux, normalized to range between -1 to +1, and colored green for positive differential flux, red for negative, and black for no significant difference. If there were any association between merged-model and individual species differential flux, we would expect to see a trend in which positive merged-model differential flux is associated with more positive individual species differential flux, and vice versa. However, the plots in Figure 4.6 do not show any such trend.

Furthermore, there was also poor correlation of the differential flux when comparing any of the three metabolic disease datasets against each other, as

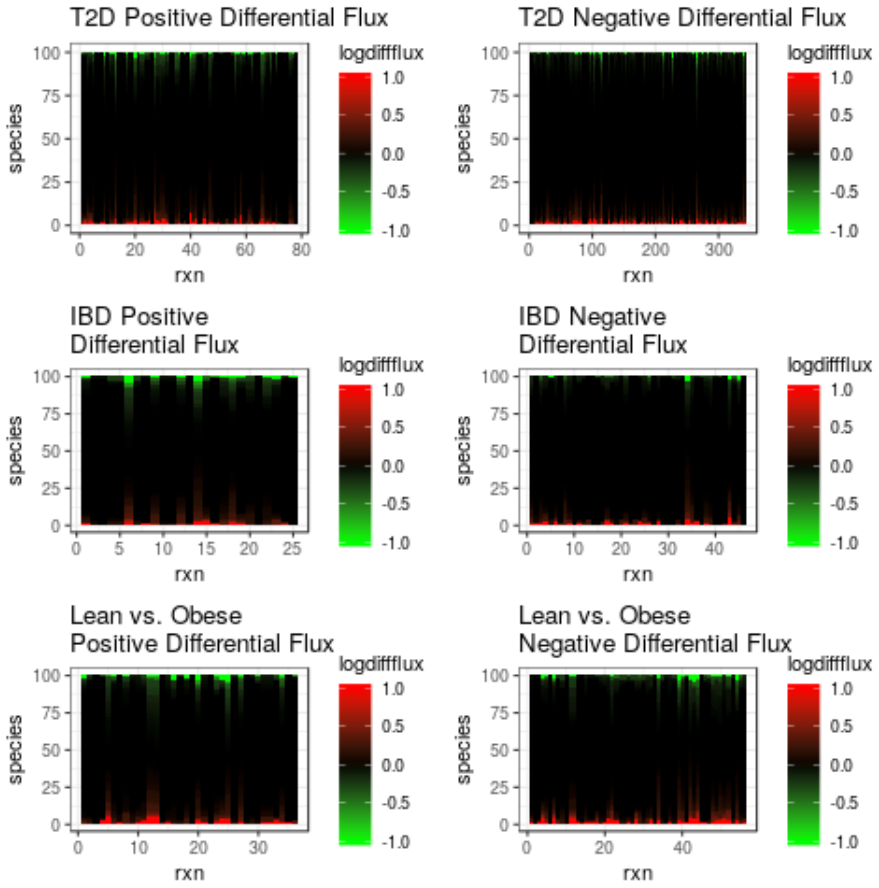


Figure 4.6: For reactions with significant merged model differential flux in each dataset, we plotted either positive (green) or negative (red) log differential flux in individual species.

shown in Tables 4.1 and 4.2. These results suggest that the overall metabolic changes between the three metabolic diseases we study are very divergent.

Finally, Figure 4.7 shows the average abundance levels and metabolic flux levels of all reactions in the three datasets plotted against each other. It shows there is very poor correlation between the level of enzyme abundance in the metagenomic data, versus the level of predicted flux. Therefore, the nature

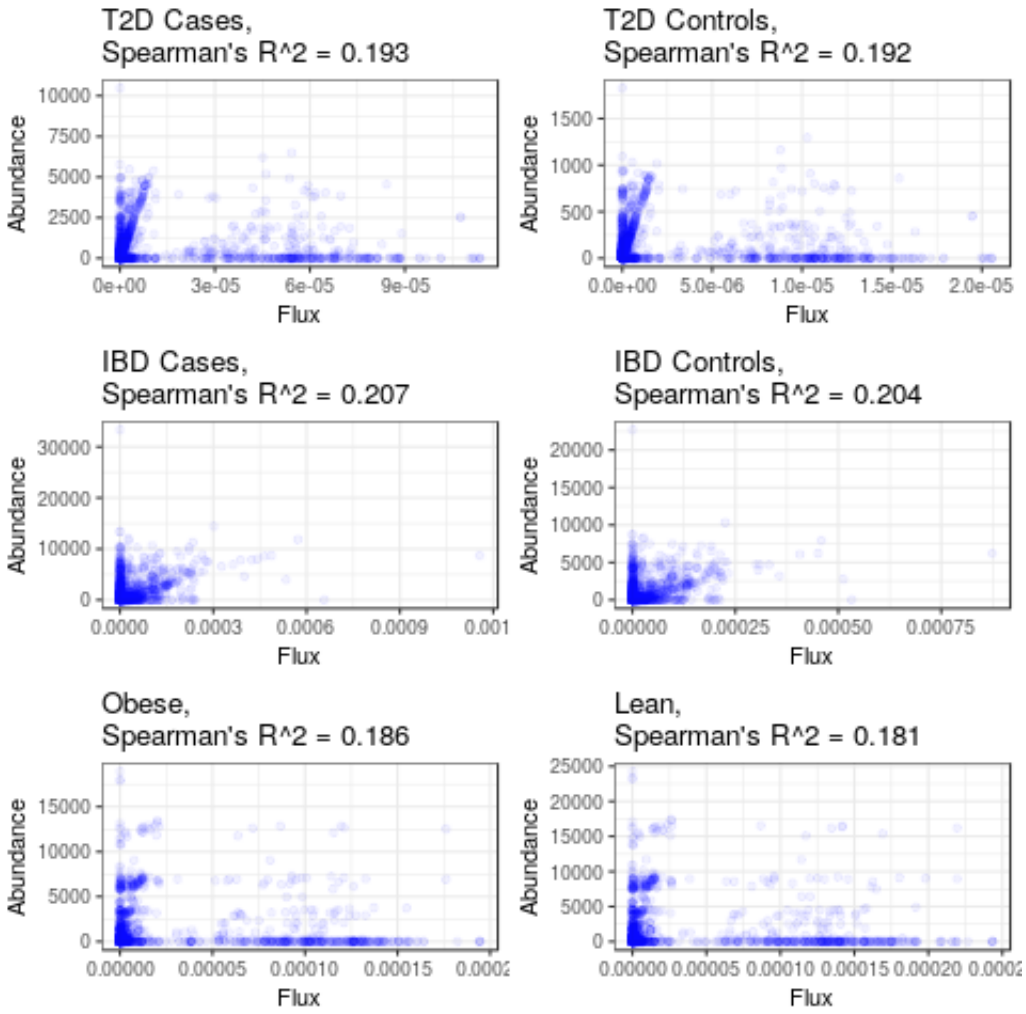


Figure 4.7: Average abundance level of reactions versus average predicted flux, across all three datasets.

of gut microbiome metabolic changes cannot be predicted from expression alone, but instead must be based on simulated flux measurements, using methods like FALCON.

Comparison Type	Spearman's R	Spearman's R-squared	Spearman's p-val
(IBD vs. normal) vs. (T2D vs. normal)	-.06	.0035	.0012
(Obese vs. normal) vs. (T2D vs. normal)	.135	.018	7.64E-12
(Obese vs. normal) vs. (IBD vs. normal)	-0.019	.00038	.33

Table 4.1: Correlations between differential flux predictions among three metabolic diseases vs. controls, for single-species models

Comparison Type	Spearman's R	Spearman's R-squared	Spearman's p-val
(IBD vs. normal) vs. (T2D vs. normal)	-.017	.00029	.39
(Obese vs. normal) vs. (T2D vs. normal)	-.026	.00068	.141
(Obese vs. normal) vs. (IBD vs. normal)	-.035	.00123	.07

Table 4.2: Correlations between differential flux predictions among three metabolic diseases vs. controls, for merged models

4.4.2 Pairwise Models

Our main interest in pairwise FALCON simulations is to measure the degree of metabolic cooperation or competition in the gut microbiome. We therefore adapted a previously published metric by Sung et al. [117] Their formula was designed to compare the number of metabolites that are both uptaken as nutrients by a pair of species, versus those metabolites that are shared by two species via crossfeeding. We modified this formula, to take into account the magnitude of fluxes involving such metabolites. Our formula for the influence of species i upon species j, through metabolic cooperation/competition, is thus

$$C_{i,j} = \frac{-\sum_k |v_{i,k}| \cdot comp_{i,j,k} + \sum_k |v_{i,k}| \cdot coop_{i,j,k}}{\sum_k |v_{i,k}|}.$$

In this equation, the index k runs over all metabolites that can be either uptaken or excreted by species i. $v_{i,k}$ represents the magnitude of the uptake or excretion flux for metabolite k in species i. Finally, the coefficient $comp_{i,j,k}$ is 1 if metabolite k is either uptaken by both i and j, or released by both of them, and 0 otherwise. Similarly, $coop_{i,j,k}$ is 1 if metabolite k is released by species i and uptaken by j, or vice versa, and 0 otherwise. The matrix C can therefore be for pairwise flux simulations involving any metagenomic sample, and represents the set of net metabolic cooperation/competition for all pairs of species.

Previous work on metabolic cooperation in the gut microbiome [117] has suggested there is a negative correlation, between the degree of cooperation in a species pair, versus the taxonomic distance between a species pair, as

measured by the Greengenes taxonomy [114]. Species pairs that are taxonomically distant are expected to differ more in their metabolic network structure and reactions. This should reduce the chance that two species both compete for the same nutrients, and also increases the probability that they crossfeed for a certain metabolite. However, surprisingly, we find that there is no correlation between the degree of metabolic cooperation and taxonomic distance among species pairs, as shown in Figure 4.8. One possible interpretation of this result is that species at all taxonomic distance actually face equal evolutionary pressure to metabolically cooperate and/or compete with each other. In that case, the magnitude of metabolic flux may be adjusted, so that overall cooperation is equal at all taxonomic distances.

From this matrix of cooperation/competition values, we then investigated the compositional stability of the gut microbiome. Compositional stability is one of the most important properties of the microbiome, as a host must have a stable microbiome in order to count on benefits from it. Our work here is inspired by a recent study from Coyte et al. [118] Their mathematical framework begins with a matrix A, where $A_{i,j}$ represents the influence of species j on growth of species i. The values of A are not known experimentally, but can be sampled from a distribution of interaction types. Positive values indicating a beneficial effect of species j's abundance on species i, and negative values a detrimental effect. The abundance and dynamics of species i, symbolized by X_i, is then given by:

$$\frac{dX_i}{dt} = X_i(r_i - s_i X_i + \sum_{j \neq i} A_{i,j} \cdot X_j)$$

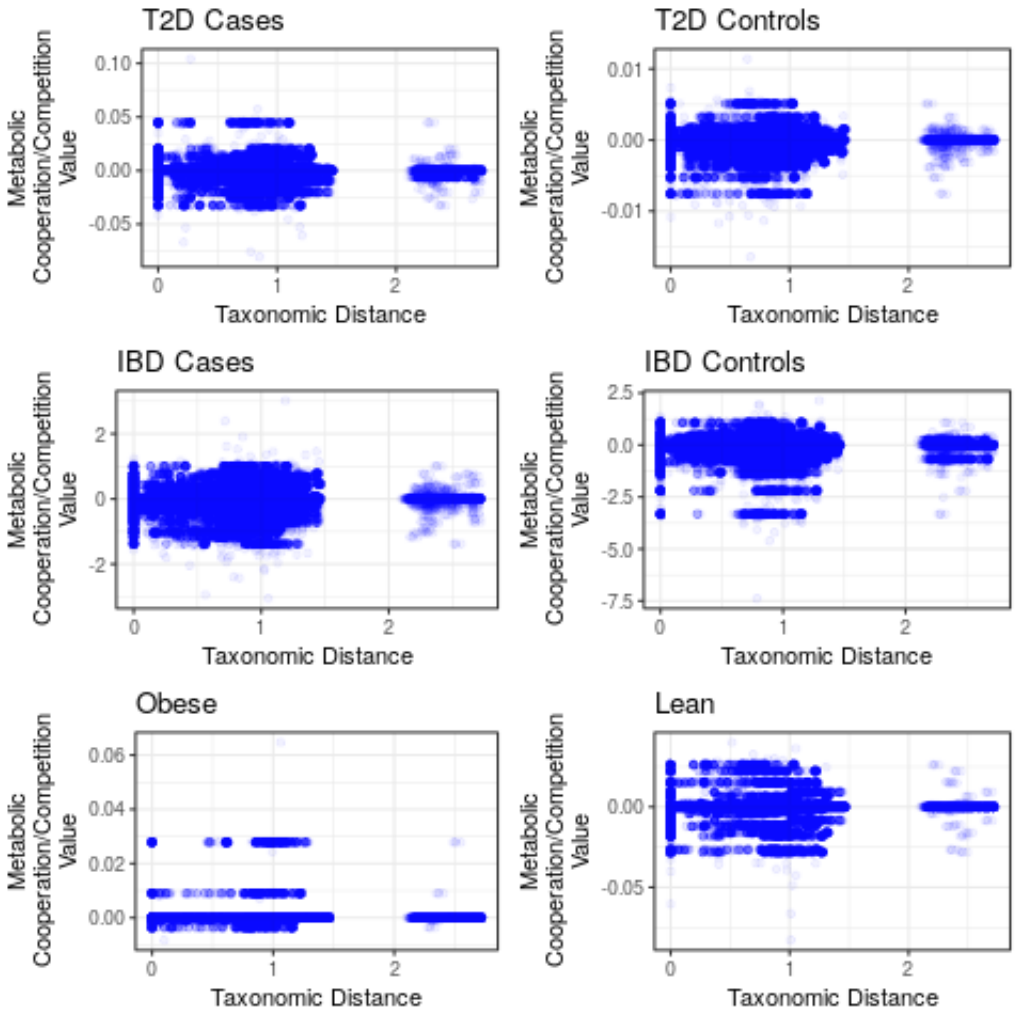


Figure 4.8: Level of metabolic cooperation vs. taxonomic distance among species pairs, in all three datasets. Note that the smaller cluster to the right of the figure represents taxonomic distances measured between archaeal and bacterial species, and the cluster to the left represents distances between two bacterial or two archaeal species. Archaeal-bacterial taxonomic distances are expected to be considerably greater than between members of the same domain.

where r_i represents the intrinsic growth rate of species i, and s_i the carrying capacity of species i. The eigenvalues of this matrix are then calculated and used to determine stability, as explained later in this paper.

The matrix C that we defined earlier has two major differences from A. Both stem from the fact that in all FALCON simulations in this study, we were unable to simulate flux through a biomass reaction, and therefore could not calculate growth rate directly. Firstly, we therefore could not determine the r_i and s_i coefficients found in the above formula, and must set them to 0 in our calculations. Secondly, whereas the coefficients of A represent the direct effects of species j on growth of species i, our matrix C represents this effect indirectly. As described before, our C contains the weighted average of metabolites under either competition or cooperation between i and j. Nevertheless, as metabolite uptake and excretion is known to be a major constraint on microbial growth rates [38], we believe our metric does capture at least some features of cooperation and competition. Therefore, we calculated cooperation/competition matrices for the datasets we described previously, and calculated eigenvalues to determine their stability.

Eigenvectors and eigenvalues are a concept from linear algebra that can be used to determine the sample's compositional stability. For any n x n square matrix, a set of n eigenvectors may be calculated, and each eigenvector is associated with one eigenvalue. The set of eigenvectors for the matrix spans all possible directions of change in the sample's composition, so that any perturbation in composition can be represented as a linear combination

of eigenvectors. Furthermore, the direction in which the composition moves, after a perturbation, is determined by the eigenvalues. Negative eigenvalues indicate that the system will move back to its original starting point after a perturbation in the direction of an eigenvector. Positive eigenvalues indicate that the system will continue moving away from the starting point, in the direction of the perturbation. Therefore, only if all eigenvalues of the cooperation/competition matrix are negative, will the composition of the system be stable.

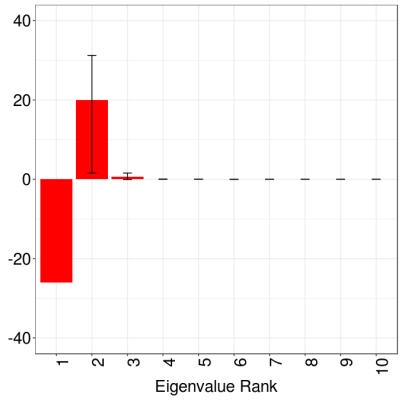
The top ten eigenvalues, by absolute value, for both normal controls and patients in all of our three datasets, are plotted in the figures below (Figure 4.9). For all of our three datasets, there were some positive eigenvalues in both normal control and patient samples, indicating that they were all unstable. However, this result is not necessarily unexpected, as all negative eigenvalues imply the **entire** microbiome is stable, so that no single species are ever lost or gained. This is actually contradicted by data such as [119], which found that a few species are lost or gained over a period of 1 year in one subject. Furthermore, we infer that the time-scale of such changes may correlate with the magnitude of the largest positive eigenvalue for each sample. From linear algebra, it is known that for a larger positive eigenvalue, a system will move more quickly away from it's original starting point, in the direction given by the corresponding eigenvector. Thus, for the T2D and IBD comparisons, the magnitude of the largest positive eigenvalue was greater in control samples compared to patients, indicating that control samples ac-

tually change more quickly than patients in these two cases. However, the opposite is true in lean samples vs obese, with lean samples having greater magnitude of the largest positive eigenvalue.

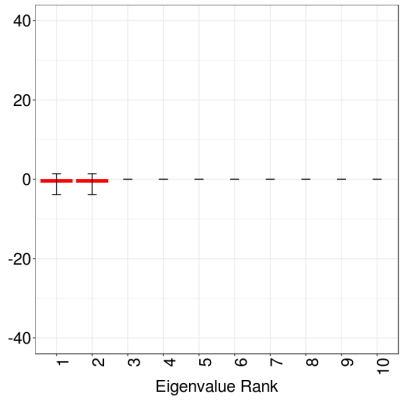
These results were unexpected for us, as previous observations suggest approximately equal stabilities of healthy and diseased microbiomes [120]. Our interpretation is that control samples in the T2D and IBD datasets do not occupy a single stable species composition, but rather change constantly within a finite space of compositions, which are consistent with having a healthy microbiome. T2D and IBD patients would also move within a finite space, corresponding to a disease state, but at a slower rate than healthy samples. Similarly, lean samples would be predicted to move within their compositional space at a slower rate than obese samples. Since 2 out of 3 metabolic diseases in this study thus are associated with slower gut microbiome compositional changes compared to controls, further studies may show whether or not this is a general trend among all microbiome-associated diseases.

4.5 Discussion

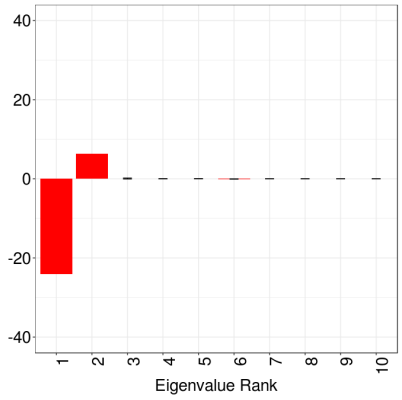
The problem of inferring metabolic flux distributions is a key problem in computational biology, and is essential to unraveling the mechanisms of many complex metabolic diseases. So far, most work in this area has focused on single bacterial species, but the gut microbiome is made up of hundreds of



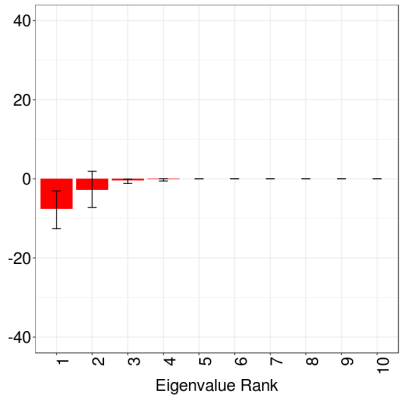
(a) Eigenvalues of normal control samples in diabetes study.



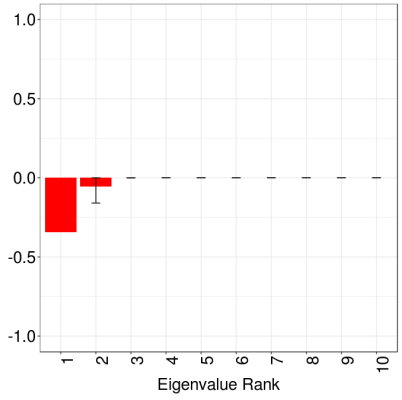
(b) Eigenvalues of type 2 diabetes samples.



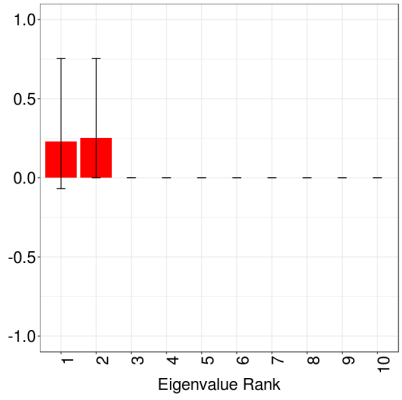
(c) Eigenvalues of normal control samples in IBD study.



(d) Eigenvalues of IBD samples.



(e) Eigenvalues of lean samples.



(f) Eigenvalues of obese samples.

Figure 4.9: Eigenvalues of pairwise models for diabetic, IBD, and obese samples, versus respective controls.

species, which together play a major role in the metabolism of the human host. To predict flux in this challenging environment, we modeled the gut microbiome with three separate models, either individual-based models, pairwise models, or a merged model, as described in Methods. We then applied a gene-expression based method, called FALCON, to infer flux distributions for metagenomic samples from either normal controls, or patients with T2D, IBD, or obesity.

Using this approach, we first identified metabolic pathways that had significant differential flux between controls and patients. Somewhat surprisingly, we observed that the overall correlation between differential metabolic flux in each of the three metabolic diseases is surprisingly low. These three metabolic diseases have high co-occurrence with each other, and are often considered to share similar metabolic causes and outcomes in the human host, and thus possibly the gut microbiome as well. However, our results suggest that, when looking at metabolism as a whole, which includes many pathways outside of central carbon metabolism, there may be little similarity between them. This may be relevant in the future, as new approaches that target novel metabolic pathways for treatment of one disease, may not apply to the others.

Nevertheless, at the level of pathways, we found that all three diseases have a few pathways with a very significant differential flux, and that a few of these are shared among these diseases. Among these is branched amino acid synthesis, which has previously been shown to be involved in human

host metabolic changes. Our results suggest that the microbiome as well may contribute to disease-specific metabolic changes, and thus may be a promising target for therapies, as several available probiotics already set out to do.

Another important question in the gut microbiome is the extent of metabolic cooperation or competition among species. Interactions among species help to determine the overall composition of the gut microbiome, as well as its stability in the face of perturbations. Here, our results were unexpected, as by using the eigenvalues of the matrix of interactions, we found that there are some positive eigenvalues in nearly all samples, indicating that all samples are unstable. Although this result is different from previous work [118], we believe there is some justification for it, based on the observation that real microbiomes do slowly gain and lose species over time [119]. Furthermore, our results show that the time-scale at which microbiome compositions change is different for metabolic disease vs. control, with T2D and IBD patients changing more quickly than controls, while obese samples change more slowly than lean.

In conclusion, our study is an attempt to model the metabolism of the gut microbiome, using gene-expression based modeling, and how it may contribute to common metabolic diseases in humans. Our results give insight into how the gut microbiome may affect common metabolic symptoms and mechanisms in these diseases. We expect that future studies may further improve the accuracy of metabolic inference in the gut microbiome.

Bibliography

[1] Cairns, R., Harris, I. and Mak, T. W. (2011). Regulation of cancer cell metabolism. Nature Reviews Cancer. *11*, 85-95

[2] Koppenol, W. H., Bounds, P. L. and Dang, C. V. (2012). Otto Warburg's contributions to current concepts of cancer metabolism. Nature Reviews Cancer. *11*, 325-337

[3] Locasale, J. W. (2013). Serine, glycine and one-carbon units: cancer metabolism in full circle. Nature Reviews Cancer. *13*, 572-583

[4] Altman, B. J., Stine, Z. E. and Dang, C. V. (2016). From Krebs to clinic: glutamine metabolism to cancer therapy. Nature Reviews Cancer. *16*, 619-634

[5] Rohrig, F. and Schulze, A. (2016). The multifaceted roles of fatty acid synthesis in cancer. Nature Reviews Cancer. *16*, 732-749

[6] Wellen, K. E. and Thompson, C. B. (2012). A two-way street: reciprocal regulation of metabolism and signalling. Nature Reviews Molecular and Cell Biology. *13*, 270-276

[7] Cairns, R. A. and Mak, T. W. (2016). The current state of cancer metabolism. Nature Reviews Cancer. *16*, 613-614

[8] Sullivan, L. B., Gui, D. Y. and Vander Heiden, M. G. (2016). Altered metabolite levels in cahcer: implications for tumour biology and cancer therapy. Nature Reviews Cancer. *16*, 680-693

[9] Kinnaird, A., Zhao, S., Wellen, K. E. and Michelakis, E. D. (2016). Metabolic control of epigenetics in cancer. Nature Reviews Cancer. *16*, 694-707

[10] Alberti, K. G. M. M., Zimmet, P. and Shaw, J. (2005). The metabolic syndromea new worldwide definition. The Lancet. *366*, 1059-1062

[11] Giovannucci, E. (2007). Metabolic syndrome, hyperinsulemia, and colon cancer: a review. The Americal Journal of Clinical Nutrition. *86*, 836-842

[12] Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. Cell. *144*, 646-674

[13] Vander Heiden, M. G., Cantley, L. C. and Thompson, C. B. (2009). Understanding the Warburg effect: the metabolic requirements of cell proliferation. Science. *324*, 1024-1033

[14] Dayton, T. L., Jacks, T. and Vander Heiden, M. G. (2016). PKM2, cancer metabolism, and the road ahead. EMBO Reports. *17*, 1721-1730

[15] Wong, K. K., Engelman, J. A. and Cantley, L. C. (2010). Targeting the PI3K signaling pathway in cancer. Current Opinion in Genetics and Development. *20*, 87-90

[16] Martini, M., De Santis, M. C., Braccini, L. and Gulluni, F. (2014). PI3K/AKT signaling pathway and cancer: an updated review. Annals of Medicine. *46*, 372-383

[17] Laplante, M. and Sabatini, D. (2012). mTOR signaling in growth control and disease. Cell. *149*, 274-293

[18] Nakazawa, M. S., Keith, B. and Simon, M. C. (2016). Oxygen availability and metabolic adaptations. Nature Reviews Cancer. *16*, 663-673

[19] Won, K. Y., Lim, S. J., Kim, G. Y., Kim, Y. W., Han, S. A. and Song, J. Y. (2012). Regulatory role of p53 in cancer metabolism via SCO2 and TIGAR in human breast cancer. Human pathology. *43*, 221-228

[20] Chen, W. et al. (2009). Direct interaction between Nrf2 and p21 Cip1/WAF1 upregulates the Nrf2mediated antioxidant response. Molecular Cell. *34*, 663673

[21] Yang, M. and Vousden, K. H. (2016). Serine and one-carbon metabolism in cancer. Nature Reviews Cancer. *16*, 650-662

[22] Losman, J. A. et al. (2013). (R)-2hydroxyglutarate is sufficient to promote leukemogenesis and its effects are reversible. Science. *339*, 16211625

[23] Xiao, M. et al. (2012). Inhibition of KGdependent histone and DNA demethylases by fumarate and succinate that are accumulated in mutations of FH and SDH tumor suppressors. Genes and Development. *26*, 13261338

[24] Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. and Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. Nature. *489*, 220230

[25] The Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human micobiome. Nature. *486*, 207-214

[26] Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T. and Pons, N. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature. *464*, 5965

[27] Shao, Y., Forster, S. C., Tsaliki, E., Vervier, K., Strang, A., Simpson, N., Kumar, N., Stares, M. D., Rodger, A., Brocklehurst, P., Field, N. and Lawley, T. (2019). Stunted microbiota and opportunistic pathogen colonization in caesarean-section birth. Nature. *574*, 117-121

[28] Sommer, F. and Backhed, F. (2013). The gut microbiota - masters of host development and physiology. Nature Reviews Microbiology. *11*, 227-238

[29] Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L. and Gordon, J. I. (2011). Human nutrition, the gut microbiome and the immune system. Nature. *474*. 327-336

[30] Koropatkin, N. M., Cameron, E. A. and Martens, E. C. (2012). How glycan metabolism shapes the human gut microbiota. Nature Reviews Microbiology. *10*, 323-335

[31] David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, A. S., Varma, Y., Fischbach, M. A., Biddinger, S. B., Dutton, R. J. and Turnbaugh, P. J. (2014). Diet rapidly and reproducibly alters the human gut microbiome. Nature. *505*, 559-563

[32] Forslund, K., Hildebrand, F., Nielsen, T., Falony, G., Le Chatelier, E., Sunagawa, S., Prifti, E. and Vieira-Silva, S. (2015). Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota. Nature. *528*, 262266

[33] Schirmer, M., Franzosa, E. A., Lloyd-Price, J., McIver, L. J., Schwager, R., Poon, T. W., Ananthakrishnan, A. N. and Andrews, E. (2018) Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. Nature Microbiology. *3*, 337346

[34] Koppel, N., Rekdal, V. M. and Balskus, E. P. (2017). Chemical transformation of xenobiotics by the human gut microbiota. Science. *356*

[35] H. J. Haiser et al., (2013). Predicting and manipulating cardiac drug inactivation by the human gut bacterium Eggerthella lenta. Science. *341*, 295298

[36] B. D. Wallace et al., (2010). Alleviating cancer drug toxicity by inhibiting a bacterial enzyme. Science. *330*, 831835

[37] Z. Wang et al., (2015). Non-lethal inhibition of gut microbial trimethylamine production for the treatment of atherosclerosis. Cell. *163*, 15851595

[38] Orth, J. D., Thiele, I. and Palsson, B. O. (2010). What is flux balance analysis? Nature Biotechnology. *28*, 245248

[39] Price, N. D., Reed, J. and Palsson, B. O. (2004). Genome-scale models of microbial cells: Evaluating the consequences of constraints. Nature Reviews Microbiology. *2*, 886-897

[40] Varma, A. and Palsson, B. (1994). Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. Nature Biotechnology. *12*, 994-998

[41] Harcombe, W., Delaney, N., Leiby, N. et al. (2013). The Ability of Flux Balance Analysis to Predict Evolution of Central Metabolism Scales with the Initial Distance to the Optimum. PLoS Computational Biology. *9*

[42] Ibarra, R., Edwards, J., and Palsson, B. (2002). Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature. *420*, 186-189

[43] Edwards, J., Ibarra, R. and Palsson, B. (2001). In silico predictions of Escherichia coli metabolic capacities are consistent with experimental data. Nature Biotechnology. *19*, 125-130

[44] Bordbar, A., Feist, A., Usaite-Black, R. et al. (2011). A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. BMC Systems Biology. *5*

[45] Chang, R. L., Xie, L., Xie, L., Bourne, P. E. and Palsson, B. O. (2010). Drug Off-Target Effects Predicted Using Structural Analysis in the Context of a Metabolic Network Model. PLoS Computational Biology. *6*

[46] Jerby, L., Shlomi, T. and Ruppin, E. (2010). Computational reconstruction of tissue-specific metabolic models application to human liver metabolism. Molecular Systems Biology. *6*

[47] Blazier, A. and Papin, J. (2012). Integration of expression data in genome-scale metabolic network reconstructions. Frontiers in Physiology. *3*

[48] Lee, D., Smallbone, K., Dunn, W., Murabito, E., Winder, C. et al. (2012). Improving metabolic flux predictions using absolute gene expression data. BMC Systems Biology. *6*

[49] Shlomi, T., Cabili, M. N., Herrgrd, M. J., Palsson, B. and Ruppin, E. (2008) Network-based prediction of human tissue-specific metabolism. Nature Biotechnology. *26*, 10031010

[50] Zur, H., Ruppin, E. and Shlomi, T. (2010). iMAT: an integrative metabolic analysis tool. Bioinformatics. *26*, 31403142

[51] Becker, S. and Palsson, B. (2008). Context-specific metabolic networks are consistent with experiments. PLoS Comp. Biol. *4*: e1000082. doi:10.1371/journal.pcbi.1000082.

[52] Colijn, C., Brandes, A., Zucker, J., Lun, D., Weiner, B. et al. (2009). Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. PLoS Computational Biology. *5*

[53] Smallbone, K. (2014). From genes to fluxes. https://u003f.wordpress.com/2014/02/19/from-genes-to-fluxes-2/

[54] Zengler, K. and Palsson, B. (2012). A roadmap for the development of community systems (CoSy) biology. Nature Reviews Microbiology. *10*, 366-372.

[55] Klunemann, M., Schmid, M. and Patil, K. (2014). Computational tools for modeling xenometabolism of the human gut. Trends in Biotechnology. *32*, 157-165.

[56] Stolyar, S., Van Dien, S., Hillesland, K., Pinel, N., Lie, T. J., Leigh, J. A. and Stahl, D. A. (2007). Metabolic modeling of a mutualistic microbial community. Molecular Systems Biology. *3*

[57] Zomorrodi, A.R. and Maranas, C.D. (2012). OptCom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. PLoS Comp. Biol. *8*, e1002363.

[58] Shoaie, S., Ghaffari, P., Kovatcheva-Datchary, P., Mardinoglu, A., Sen, P., Pujos-Guillot, E., de Wouters, T., Juste, C. et al. (2015). Quantifying Diet-Induced Metabolic Changes of the Human Gut Microbiome. Cell Metabolism. *22*, 320-31.

[59] Chan, S. H. J., Simons, M. N. and Maranas, C. D. (2017). SteadyCom: predicting microbial abundances while ensuring community stability. PLoS Comput. Biol. *13*, e1005539.

[60] Garza, D. R., van Verk, M. C., Huynen, M. A. and Dutilh, B. E. (2018). Towards predicting the environmental metabolome from metagenomics with a mechanistic model. Nature Microbiology. *3*, 456-460.

[61] Kumar, M., Ji, B., Zengler, K. and Nielsen, J. (2019). Modelling approaches for studying the microbiome. Nature Microbiology. *4*, 1253-1267.

[62] Gibbons, C., Montgomery, M., Leslie, A., and Walker, J. (2000). The structure of the central stalk in bovine F1-ATPase at 2.4 A resolution. Nat. Struct. Mol. Biol., *7*, 10551061

[63] Karr, J., Sanghvi, J., Macklin, D., et al. (2012). A whole-cell computational model predicts phenotype from genotype. Cell. *150*, 389401

[64] Mahadevan, R., Edwards, J.S. and Doyle, F.J. (2002). Dynamic flux balance analysis of diauxic growth in Escherichia coli. Biophysical Journal. *83*, 1331-1340

[65] Barker, B. E., Sadagopan, N., Wang, Y. and Smallbone, K. (2015). A robust and efficient method for estimating enzyme complex abundance and metabolic flux from expression data. Computational Biology and Chemistry. *59*, 98-112

[66] Jain, M., Nilsson, R., Sharma, S. et al. (2012). Metabolite Profiling Identifies a Key Role for Glycine in Rapid Cancer Cell Proliferation. Science. *336*, 1040-1044

[67] Machado, D. and Herrgard, M. (2014). Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism. PLoS Computational Biology. *10*

[68] Gholami, A., Hahne, H., Wu, Z. et al. (2013). Global Proteome Analysis of the NCI-60 Cell Line Panel. Cell Reports. *4*, 609620

[69] Feist, A., Herrgard, M., Thiele, I, and Palsson, B. (2009). Reconstruction of biochemical networks in microorganisms. Nature Reviews Microbiology. *7*, 129-143

[70] Gudmundsson, S. and Thiele, I. (2010). Computationally efficient flux variability analysis. BMC Bioinformatics. *11*

[71] Fruman, D. A., Chiu, H., Hopkins, B. D., Bagrodia, S., Cantley, L. C. and Abraham, R. (2017). The PI3K Pathway in Human Disease. Cell. *170*, 605-635

[72] Dang, C. V. (2012). MYC on the path to cancer. Cell. *149*, 22-35

[73] Wise, D. R., DeBerardinis, R. J., Mancuso, A., Sayed, N., Zhang, X.-Y., Pfeiffer, H. K., Nissim, I., Daikhin, E., Yudkoff, M., McMahon, S. B. and Thompson, C. B. (2008). Myc regulates a transcriptional program that stimulates mitochondrial glutaminolysis and leads to glutamine addiction. Proc. Nat. Acad. of Sciences. *105*, 18782-18787

[74] Sauer, U. (2006). Metabolic networks in motion: 13Cbased flux analysis. Molecular Systems Biology.*2*

[75] Antoniewicz, M. R. (2015). Methods and advances in metabolic flux analysis: a mini-review. Journal of Industrial Microbiology and Biotechnology. *42*, 317325

[76] The Cancer Genome Atlas Research Network (TCGARN). http://cancergenomie.nih.gov/.

[77] Thiele, I., Swainston, N., Fleming, R. M. T., Hoppe, A., Sahoo, S., Aurich, M. K., Haraldsdottir, H. and Mo, M. L. (2013). A community-driven global reconstruction of human metabolism. Nature Biotechnology. *31*, 419425

[78] Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., Shen, R. and Taylor, A. M. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. *173*, 291-304

[79] Machado, D. and Herrgrd, M. (2014). Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. PloS Computational Biology. *10*

[80] Kochanowski, K., Sauer, U. and Chubukov, V. (2013). Somewhat in control the role of transcription in regulating microbial metabolic fluxes. Current Opinion in Biotechnology. *24*, 987993

[81] Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A. and Wendl, M. C. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. *173*, 371-385

[82] Gaude, E. and Frezza, C. (2016). Tissue-specific and convergent metabolic transformation of cancer correlates with metastatic potential and patient survival. Nature Communications. *7*

[83] Subramanian, A. 1., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A. and Pomeroy, S. L. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Nat. Acad. of Sciences. *102*, 15545-15550

[84] Gao, X., Sanderson, S. M., Dai, Z., Reid, M. A., Cooper, D. E., Lu, M., Richie Jr, J. and Ciccarella, A. (2019). Dietary methionine influences therapy in mouse cancer models and alters human metabolism. Nature. *572*, 397401

[85] Ogretmen, B (2018). Sphingolipid metabolism in cancer signalling and therapy. Nature Reviews Cancer. *18*, 33-50

[86] Ma, C., Han, M., Heinrich, B., Fu, Q., Zhang, Q., Sandhu, M., Agdashian, D. and Terabe, M. (2018). Gut microbiomemediated bile acid metabolism regulates liver cancer via NKT cells. Science. *360*

[87] Finicle, B. T., Jayashankar, V and Edinger, A. L. (2018). Nutrient scavenging in cancer. Nature Reviews Cancer. *18*, 619633

[88] Ding, X., Zhang, W., Li, S. and Yang, H. (2019). The role of cholesterol metabolism in cancer. The American Journal of Cancer Research. *9*, 219227

[89] Pinho, S. S. and Reis, C. A. (2015). Glycosylation in cancer: mechanisms and clinical implications. Nature Reviews Cancer. *15*, 540-555

[90] Israelsen, W. J., Dayton, T. L., Davidson, S. M., Fiske, B. P., Hosios, A. M., Bellinger, G., Li, J., Yu, Y. and Sasaki, M. (2013). PKM2 isoform-specific deletion reveals a differential requirement for pyruvate kinase in tumor cells. Cell. *155*, 397-409

[91] Mamede, A. C., Tavares, S. D., Abrantes, A. M., Trindade, J., Maia, J. M. and Botelho, M. F. (2011). The Role of Vitamins in Cancer: A Review. Nutrition and Cancer. *63*, 479-494

[92] Gianchandani, E. P., Chavali, A. K. and Papin, J. A. (2010). The application of flux balance analysis in systems biology. Wiley Interdisciplinary Reviews: Systems Biology and Medicine. *2*, 372-382

[93] Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J. and Palsson, B. O. (2004). Integrating high-throughput and computational data elucidates bacterial networks. Nature. *429*, 92-96

[94] Ward, P. S. and Thompson, C. B. (2012). Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. Cancer Cell. *21*, 297308

[95] Musso, G., Gambino, R. and Cassader, M. (2011). Interactions between gut microbiota and host metabolism predisposing to obesity and diabetes. Annual Review of Medicine. *62*, 361-380

[96] Lakka, H. M., Laaksonen, D. E., Lakka, T. A., Niskanen, L. K., Kumpusalo, E., Tuomilehto, J. and Salonen, J. T. (2002). The metabolic syndrome and total and cardiovascular disease mortality in middle-aged men. The Journal of the American Medical Association. *288*, 2709-2716

[97] Cho, I. and Blaser, M. (2012). The human microbiome: at the interface of health and disease. Nature Reviews Genetics. *13*, 260270

[98] Tremaroli, V. and Backhed, F. (2012). Functional interactions between the gut microbiota and host metabolism. Nature. *489*, 242-249

[99] Goodrich, J. K., Davenport, E. R. and Clark, A. G. (2017). The relationship between the human genome and microbiome comes into view. Annual Review of Genetics. *51*, 413-433

[100] Wang, Z., Klipfell, E., Bennett, B. J., Koeth, R., Levison, B. S., DuGar, B., Feldstein, A. E. Britt, B. E., Fu, X. et al. (2011). Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. Nature. *472*, 5763

[101] Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S. and Zhang, W. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. *490*, 5560

[102] den Besten, G., van Eunen, K., Groen, A. K., Venema, K., Reijngoud, D. J. and Bakker, B. M. (2013). The role of short-chain fatty acids in the interplay between diet, gut microbiota, and host energy metabolism. The Journal of Lipid Research. *54*, 2325-2340

[103] Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. Nature. *486*, 207-214

[104] Quek, L. E., Dietmair, S., Krmer, J. O. and Nielsen, L. K. (2010). Metabolic flux analysis in mammalian cell culture. Metabolic Engineering. *12*, 161-171

[105] Heinken, A., Sahoo, S., Fleming, R. M. and Thiele, I. (2013). Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. Gut Microbes. *4*, 28-40

[106] Shoaie, S., Ghaffari, P., Kovatcheva-Datchary, P., Mardinoglu, A., Sen, P., Pujos-Guillot, E., de Wouters and T. and Juste, C. (2015). Quantifying Diet-Induced Metabolic Changes of the Human Gut Microbiome. Cell Metabolism. *22*, 320-331

[107] Magnsdttir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., Greenhalgh, K. and Jger, C. (2017). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nature Biotechnology. *35*, 8189

[108] Foster, K. R., Schluter, J., Coyte, K. Z. and Rakoff-Nahoum, S. (2017). The evolution of the host microbiome as an ecosystem on a leash. Nature. *548*, 4351

[109] Kopylova, E., No, L. and Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. *28*, 3211-3217

[110] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glckner, F. O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research. *41*, D590D596

[111] Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H. and Alm, E. J. (2018). QIIME

2: Reproducible, interactive, scalable, and extensible microbiome data science. Nature Biotechnology. *37*, 852857

[112] DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D. and Hu, P. (2016). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology. *72*, 5069-5072

[113] Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., Clemente, J. C. and Burkepile, D. E. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature Biotechnology. *31*, 814821

[114] McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A, Andersen, G. L. and Knight, R. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. The ISME Journal. *6*, 610618

[115] Hooper, L. V., Midtvedt, T. and Gordon, J. I. (2002). How host-microbial interactions shape the nutrient environment of the mammalian intestine. Annual Review of Nutrition. *22*, 283-307

[116] Tailford, L. E., Crost, E. H., Kavanaugh, D. and Juge, N. (2015). Mucin glycan foraging in the human gut microbiome. Frontiers in Genetics. *6*

[117] Sung, J., Kim, S., Cabatbat, J. J. T., Jang, S., Jin, Y-S., Jung, G. Y., Chia, N. and Kim, P-J. (2017). Global metabolic interaction network of the human gut microbiota for context-specific community-scale analysis. Nature Communications. *8*

[118] Coyte, K. Z., Schluter, J. and Foster, K. R. (2015). The ecology of the microbiome: networks, competition, and stability. Science. *350*, 663-666

[119] Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., Knights, D. and Gajer, P. (2011). Moving pictures of the human microbiome. Genome Biology. *12*

[120] Stein, R. R., Bucci, V., Toussaint, N. C., Buffie, C. G., Rtsch, G., Pamer, E. G., Sander, C. and Xavier, J. B. (2013). Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota. PloS Computational Biology. *9*